L'Homme face à son environnement : une histoire génétique et épigénétique du génome humain

Maud Fagny

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité : Génétique et épigénétique humaine

École doctorale : « Complexité du Vivant (ED 515) »

présentée par

Maud FAGNY

pour obtenir le grade de :

DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

L'Homme face à son environnement: une histoire génétique et épigénétique du génome humain

soutenue le 29 juin 2014, devant le jury composé de :

M.	**Vincent COLOT**	**Président du jury**
M.	**Ludovic ORLANDO**	**Rapporteur**
M^me^	**Maud TENAILLON**	**Rapporteur**
M.	**Guillaume ACHAZ**	**Examinateur**
M.	**Laurent EXCOFFIER**	**Examinateur**
M.	**Lluís QUINTANA-MURCI**	**Directeur de thèse**

A Mémé, à Mamie

Remerciements

Tout d'abord, je remercie tous les membres de mon jury : Vincent Colot, qui m'a fait l'honneur d'accepter de le présider, Guillaume Achaz et Laurent Excoffier, les examinateurs. Merci à Maud Tenaillon et Ludovic Orlando pour avoir accepté de d'être les rapporteurs de cette thèse. Ludovic, merci beaucoup de m'avoir orienté vers le labo de Lluis pour mon deuxième stage de M2, l'existence de cette thèse doit beaucoup à ce conseil.

Lluis, merci de m'avoir fait confiance et de m'avoir accueillie dans ton laboratoire pour cette thèse. Les discussions avec toi tout au long de cette thèse, et en particulier pendant ces fameux PhDmeetings ont été très enrichissantes et cruciales pour l'avancement de mes travaux de recherche et la réalisation de cette thèse. Merci infiniment pour tous les conseils, les relectures et les encouragements. Je n'aurais pas pu espérer un meilleur encadrement.

Je remercie également Guillaume qui a joué un rôle prépondérant pendant ma thèse. Merci d'avoir accepté de m'encadrer pendant mon stage de M2. J'ai été ravie de travailler avec toi. Nos échanges ont toujours été très intéressant, qu'il s'agisse de science ou de politique.

Etienne, merci pour les conseils et les discussions scientifiques toujours très intéressantes. Maxime, bien que tu ne sois arrivé au labo que depuis 10 mois, tes conseils en statistiques et analyses de résultats ont été très utiles dans la dernière ligne droite. Merci à Hélène, Christine et Nora pour toutes les manip'. Sans vous, je n'aurais rien eu à analyser comme données pendant ma thèse.

Merci à tous les membres du labo, ceux qui y sont aujourd'hui et ceux qui en sont parti. Hélène, merci pour tous tes efforts culinaires, les cookies, le thé glacé, les barbecues et plus généralement pour tout le travail que tu fais en tant que labmanager. Ta présence et tes efforts contribue énormément à la cohésion du groupe. Christine, merci pour ta présence au quotidien. Ton soucis permanent des autres et le sérieux avec lequel tu rempli ton rôle de correspondante hygiène et sécurité font tenir debout le labo. Katie, ma voisine de bureau de ces deux dernières années et ma colocataire pour les conférences, merci. Ces années de thèses n'auraient pas été aussi fun sans toi. Merci à Marie et Julien, vos conseils avisés sur tous styles de musique ont été particulièrement utiles pour renouveler ma playlist. Merci à Hélène, Nora, Maxime et Matthieu, le noyau dur du groupe déjeuner, vous avez contribué à faire des pauses déjeuner de vrais moments de détente. Plus généralement, merci à toutes les thésards et stagiaires de l'équipe, Jérémy, Katie, Matthieu, Marie, Eric, Lucas, Stéphane et Choumouss pour la bonne humeur. Merci également à Eddie et Barbara pour vos coups de main discrets mais efficaces. Enfin, merci à Cécile pour ta gestion au quotidien toutes les questions administratives.

Un grand merci à toute l'équipe de volley de pasteur pour tous les bons moments passés ensembles, les matchs (les rares gagnés et les nombreux perdus) et les pizzas du mercredi soir. Ces moments auront contribué à m'aider à évacuer mon stress et à

contrebalancer (un peu) l'effet de la cantine de Pasteur.

Merci à toute la bande des amis. Merci aux rascasses pour tous les bons moments passés ensembles, les séjours à Condors, les discussions sans queue ni tête jusqu'à pas d'heure durant lesquelles des théories scientifiques pour le moins intéressantes auront été développées... Merci en particulier à Phil, sans qui je ne serait probablement pas allée aussi souvent au théâtre ces quatre dernières années, et à Manon, pour tous les bons moments passés ensembles à l'ENS et après. C'est bientôt à ton tour de soutenir, et je te souhaite plein de courage d'ici là. Je suis contente de repartir explorer un bout du monde avec vous deux.

Ma thèse doit également beaucoup à Sylvain Mousset, qui m'a appris à programmé et m'a offert l'opportunité de découvrir la recherche en génétique des populations. Sylvain et Ludovic, vos cours à l'ENS de Lyon étaient parmi les meilleurs auxquels j'ai assisté et ont beaucoup contribué à éveiller mon intérêt pour la génétique des populations et l'évolution.

Merci également à mes parents. Vous m'avez donné envie d'apprendre sans cesse et de découvrir de nouveaux horizons. Vous m'avez toujours soutenu et avez su créez les conditions de mon épanouissement personnel. Merci à vous et à Laure pour toute votre affection et votre patience, vos encouragements et votre soutient tout au long de cette aventure, il a été particulièrement important. Et non moins important pour l'écriture de cette thèse, merci pour le travail de relecture.

Enfin, Pierre, merci infiniment pour ton soutient au cours de cette thèse, particulièrement dans les moments les plus difficiles. Merci aussi pour les coups de main en programmation. Mais surtout merci pour toutes les petites choses qui seraient trop longues à citer, mais qui rendent ta présence quotidienne particulièrement précieuse.

d

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

CLR *Composite likelihood ratio*

CMS *Composite of multiple signals*

iHS *Integrated haplotype score*

5-caC 5-carboxylcytosine

5-fC 5-formylcytosine

5-hmC 5-hydroxyméthylcytosine

5-mC 5-méthylcytosine

A Adénosine

ADN Acide désoxyribonucléique

AGR *Agriculturalists*, populations d'agriculteurs d'Afrique Centrale.

ARN Acide ribonucléique

ARNnc Longs ARN non codants

ARNpi Petits ARN interagissant avec les protéines Piwi

C Cytosine

CNV *Copy number variation*, type de variation structurelle correspondant à une répétition d'un fragment de chromosome

CpA Dinucléotide cytosine-adénosine

CpG Dinucléotide cytosine - guanine

CpT Dinucléotide cytosine-thymine

DIND *Derived intra-allelic nucleotide diversity*

DMS *Differentially methylated sites*, sites différentiellement méthylés

DNMT ADN méthyltransférase

EHH *Extended haplotype homozygosity*

FDR *False discovery rate*, taux de fausses découvertes

G Guanine

GWAS *Genome-wide association studies*, études d'association entre SNP et phénotypes à l'échelle du génome

H3K27 Méthylation du résidu lysine 27 de l'histone H3

LRH *Long range haplotype*

LRT *Likelihood ratio test*

LSBL *Locus-specific branch lengths*

MBD *Methyl-CpG-binding domain*, famille de protéines possédant un domaine capable de lier les 5-mCpG

meQTL *Methylation quantitative trait loci*, mutation génétique associée à des variations du niveau de méthylation

PBS *Population branch statistics*

RHG *Rainforest hunter-gatherers*, populations de chasseurs-cueilleurs vivant dans la forêt équatoriale.

SNP *Single nucleotide polymorphism*, substitution d'un nucléotide de la séquence d'ADN par un autre

T Thymine

TLR *Toll-like receptors*, récepteurs de type Toll

UV Ultraviolets

INTRODUCTION

La diversité phénotypique observée au sein de l'espèce humaine provient essentiellement de deux sources : la diversité génétique et la diversité épigénétique (figure 1). Plusieurs autres forces participent à forger cette diversité, dont l'environnement, qui exerce une influence à la fois à long terme sur la variabilité génétique et à court terme sur la variabilité épigénétique. Au cours de son histoire, l'Homme a connu de nombreux bouleversements de son environnement, certains involontaires, comme la fin de la dernière ère glaciaire. De par la diversité de ses cultures et de ses sociétés, ainsi que de ses nombreuses inventions techniques, l'espèce humaine a également contribué à modifier son environnement. C'est ainsi le cas de la transformation des paysages qui a suivi l'expansion de l'agriculture, puis beaucoup plus récemment de la très forte augmentation de l'urbanisation qui a découlé de la révolution industrielle. Ces modifications volontaires du paysage ont souvent été accompagnées d'autres conséquences environnementales moins volontaires mais pouvant avoir tout autant d'impact sur la diversité humaine. C'est ainsi le cas de la transformation de l'environnement pathogénique pendant le Néolithique. Celle ci s'est traduite à la fois par une exposition à de nouveaux agents pathogènes provoquée par la promiscuité résultant de la domestication des animaux et par l'émergence de maladies infectieuses épidémiques dont la transmission a été facilitée par l'augmentation de la densité des populations. La multiplication des échanges commerciaux et le réchauffement climatique accéléré par l'activité humaine sont aujourd'hui la source de changements environnementaux majeurs avec notamment l'apparition ou la réapparition de certains pathogènes dans diverses régions de la planète.

Bien que la diversité génétique ait été largement étudiée depuis la seconde moitié du XXème siècle, le séquençage du génome humain, il y a 14 ans, a permis d'envisager son étude à l'échelle du génome entier. Des débats ont alors été réactivés, notamment sur l'importance de l'action de l'environnement sur l'évolution humaine via la sélection naturelle à l'échelle de plusieurs générations. Enfin, ces dernières années, l'importance de la diversité épigénétique sur la diversité phénotypique a été soulignée, et il a été montré que des facteurs environnementaux peuvent avoir un impact immédiat sur le profil épigénétique d'un individu. Durant ma thèse, je me suis donc intéressée à l'étude de l'impact de l'environnement, à différentes échelles de temps, sur la diversité génétique et épigénétique humaine. Dans un premier temps, j'ai

évalué l'existence et l'importance d'une classe particulière d'événements de sélection naturelle, les balayages sélectifs, dans l'évolution « récente » de la diversité génétique humaine (à une échelle d'environ 30 000 ans). Dans un deuxième temps, je me suis concentrée sur l'étude de l'impact respectif des facteurs génétiques et des différences environnementales passées et actuelles sur la diversité épigénétique des populations humaines.

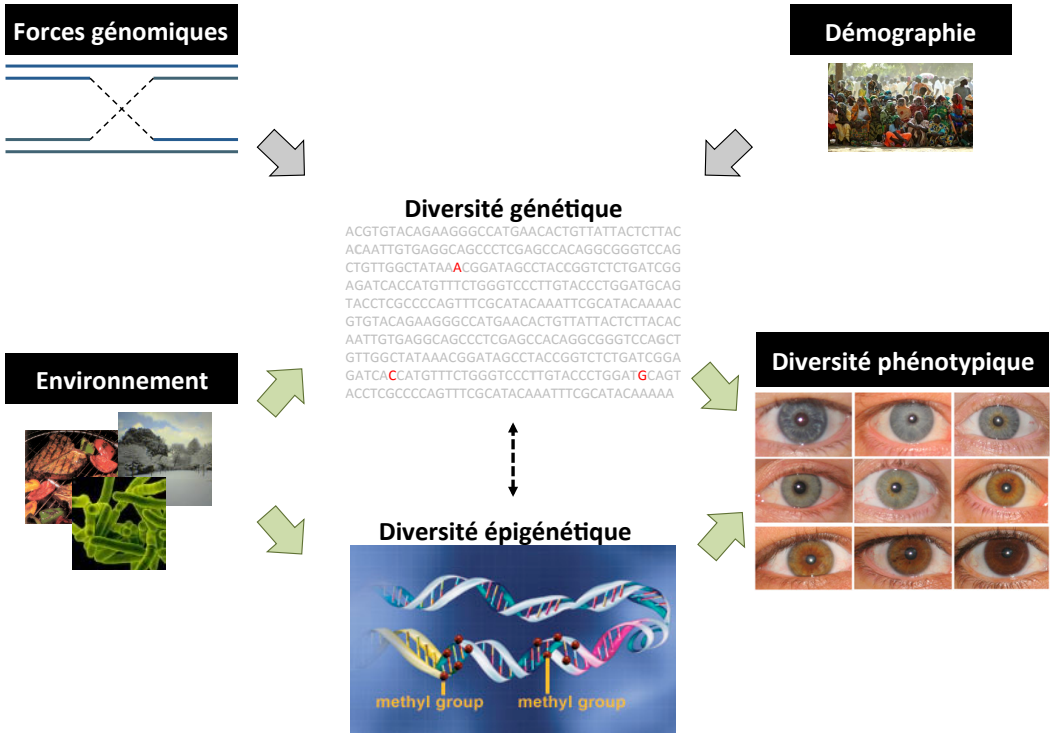


Fig. 1 Les différentes forces agissant sur la diversité phénotypique.
Les flèches pleines représentent des relations de causalité, la flèche en pointillé des relations d'association.

Chapitre 1

De la diversité génétique à la variabilité phénotypique

Dans l'espèce humaine, l'ADN (acide désoxyribonucléique) constitue le support stable de l'information génétique. Le génome humain est composé d'environ 3,1 milliards de nucléotides, et comprend entre 19,500 et 20,500 gènes codant des protéines (Clamp et al. 2007, Consortium 2001, Cunningham et al. 2014, Ezkurdia et al. 2014, International Human Genome Sequencing Consortium 2004, Venter et al. 2001) et un peu moins de 25 000 gènes non codants (Cunningham et al. 2014). Si les régions géniques constituent jusqu'à 40% du génome, seul 1% du génome, les exons, code effectivement des protéines (ENCODE Project Consortium 2012). De récentes études ont cependant montré que jusqu'à 80% du génome humain est biochimiquement fonctionnel, qu'il soit constitué de régions géniques, régulatrices de l'expression des gènes, transcrites en ARN non codants, ou associées à la structure de la chromatine (Dunham et al. 2012, Kellis et al. 2014). Au sein de l'espèce humaine, chaque individu, à l'exception des jumeaux monozygotes, possède un génome qui lui est unique, identique dans toutes ses cellules somatiques, et qui va permettre de déterminer un certain nombre de traits communs à l'ensemble de l'espèce humaine (bipédie, plan d'organisation du corps) ou spécifiques de l'individu (couleur et forme des yeux et des cheveux). Si l'on considère deux individus pris au hasard, on observe environ un SNP toutes les 1 000 pb (Consortium 2001, The 1000 Genomes Project 2010, Venter et al. 2001). A titre de comparaison, notre génome diffère de celui du chimpanzé, espèce vivante la plus proche dans l'arbre du vivant, pour entre 1 et 4% des sites (Chimpanzee Sequencing and Analysis Consortium 2005, Varki and Altheide 2005). On estime par ailleurs qu'un individu possède dans son génome de 10 000 à 11

000 mutations non synonymes qui changent la séquence d'acide aminé des protéines, sans forcément toujours avoir un impact fonctionnel, entre 340 et 400 mutations entraînant de potentielles pertes de fonctions pour 250 à 300 gènes, et est hétérozygote pour 50 à 100 mutations entraînant des maladies mendéliennes (The 1000 Genomes Project 2010). Cette diversité génétique, issue de 200 000 ans d'évolution, est donc une des sources de la variabilité phénotypique observée dans l'espèce humaine. Elle résulte de l'action de différentes forces : les forces génomiques, qui créent de la diversité, la démographie, la dérive génétique, les migrations et la sélection naturelle qui modifient les fréquences des variations génétiques au sein des populations.

1.1 Les facteurs génomiques à l'origine de la diversité génétique

1.1.1 Les polymorphismes génétiques

L'ADN peut subir différents types de modifications de sa séquence nucléotidique. On distingue les substitutions d'un acide nucléique par un autre ou SNP (pour *single nucleotide polymorphism*) des autres polymorphismes, appelés variantions structurelles. Ces derniers rassemblent les inversions de séquences plus ou moins longues, les insertions et délétions d'un ou plusieurs nucléotides, et l'amplification d'un motif répété de nucléotides (Feuk et al. 2006). Ce dernier type de polymorphisme est réparti en trois classes d'éléments en fonction de la longueur des motifs répétés. On distingue les microsatellites, de 1 à 6 pb (paires de bases), les minisatellites, de 6 à 100 pb, et les variations du nombre de copies (CNV, *copy number variation*), de quelques centaines de bases à plusieurs mégabases (Kidd et al. 2008, Redon et al. 2006, Sebat et al. 2004). Ces mutations créent de la diversité, mais ne seront transmises à la descendance que si elles apparaissent dans les cellules de la lignée germinale, à l'origine des gamètes, et pourront alors se propager dans la population.

Les SNP sont les mutations les plus nombreuses dans le génome humain et les plus étudiées en génétique des populations, notamment dans le cadre des études d'association entre génotype et phénotype (GWAS, *genome-wide association studies*) et de la détection de la sélection naturelle. Les études d'association ont permis de montrer que de nombreuses mutations pouvaient expliquer, au moins partiellement, des variations phénotypiques (Welter et al. 2014). La dernière version de la base

de données dbSNP (build 142, octobre 2014, Sherry et al. (2001)) recense un peu plus de 88 millions de SNP validés. Ces mutations proviennent principalement de deux types de mécanismes : une erreur de la machinerie de réplication de l'ADN entraînant l'insertion d'un nucléotide erroné, ou une erreur du système de réparation de l'ADN suite à une altération physique (radiation ionisante, rayons ultraviolets ou UV), ou chimique (exposition à des molécules mutagènes) de la molécule d'ADN (Friedberg 2003). Le taux de mutation dans l'espèce humaine est d'environ 10^{-8} par site et par génération. Cependant, ce taux est très variable selon les régions du génome (Campbell et al. 2012). En particulier, les dinucléotides CpG présentent un taux de mutation environ dix fois supérieur (Nachman and Crowell 2000), et certaines régions hypervariables du génomes peuvent présenter un taux de mutation allant jusqu'à 10^{-5}. On constate que les transitions, c'est-à-dire les remplacements d'une purine par une autre (adénosine (A) <-> guanine (G)) ou d'une pyrimidine par une autre (cytosine (C) <-> thymine (T)) sont beaucoup plus fréquentes que les transversions (Nei 1987). Des études récentes comparant les mutations rares spécifiques de chaque population suggèrent que ce taux de mutation pourrait avoir varié au cours de l'histoire de l'espèce humaine, avec notamment une augmentation de la fréquence de mutations spécifiques de l'exposition aux UV chez les Européens (Harris 2015).

Si les SNP sont beaucoup plus fréquents que les variations structurelles, ces derniers, de par leur longueur, représentent une part importante de la diversité génétique humaine. Ils pourraient en effet représenter jusqu'à 13% du génome (Stankiewicz and Lupski 2010). dbVar (www.ncbi.nlm.nih.gov/dbvar) recense près de 3,5 millions de variations structurelles, dont 1 868 inversions, qui peuvent concerner quelques bases ou des fragments importants de chromosomes. Les insertions ou délétions de fragments d'ADN sont aussi de tailles très variables. Le *1,000 Genomes Project* a permis d'en identifier près de 1,5 million, et dbVar compte aujourd'hui plus de 1,2 millions d'insertions. Quant aux amplifications de motifs nucléotidiques, on a découvert environ 379 000 microsatellites (microsatDB, http ://discovery.vbi.vt.edu/MicrosatDB), des centaines de types de minisatellites, chaque type pouvant être répété sur des centaines ou milliers de paires de bases (Vergnaud and Denoeud 2000), et plus de 353 000 CNV (The Database of Genomic Variants, MacDonald et al. (2014)). Ces variations structurelles proviennent d'erreurs du mécanisme de réplication, de la réparation des cassures double-brins de l'ADN,

et de la recombinaison méiotique (Conrad et al. 2010, Gu et al. 2008). Ils présentent un taux de mutation beaucoup plus élevé que les substitutions, allant de 10^{-3} à 10^{-5} (Lupski 2007) et peuvent avoir des effets importants sur le phénotype (Conrad et al. 2010, Cooper et al. 2007, Hurles et al. 2008, Stankiewicz and Lupski 2010).

1.1.2 La recombinaison méiotique

La recombinaison méiotique correspond à l'échange de matériel génétique entre deux chromosomes homologues au cours de la gamétogenèse. Une nouvelle mutation apparaît sur un haplotype donné, qui est défini comme la combinaison des allèles présents aux SNP situés sur le même chromosome. En l'absence de recombinaison, une mutation qui apparaît sur un haplotype serait systématiquement transmise à la descendance en même temps que les autres allèles déjà présents sur ce même haplotype. La recombinaison permet la création au cours des générations de nouveaux haplotypes, appelés haplotypes recombinants, et donc de nouvelles combinaisons d'allèles, permettant de casser le déséquilibre de liaison (DL) entre cette mutation et les autres.

Le taux de recombinaison, comme le taux de substitution, est extrêmement variable le long du génome humain. Il existe en effet des points chauds où le taux de recombinaison peut être supérieur de plusieurs ordres de grandeur par rapport aux régions voisines (McVean et al. 2004). On trouve environ 33 000 points chauds de recombinaison sur l'ensemble du génome, soit environ un tous les 100 kb (Myers et al. 2005, The International HapMap Consortium 2005, The International HapMap Consortium 2007). De façon intéressante, on constate que les régions géniques présentent un faible taux de recombinaison, alors que les séquences régulatrices en amont des gènes sont enrichies en points chauds de recombinaison (Kong et al. 2002, The 1000 Genomes Project Consortium 2012, The International HapMap Consortium 2007). La présence de ces points chauds dépend de plusieurs facteurs, notamment de l'éloignement par rapport aux centromères, de la richesse en nucléotides C et G de la région, et de la présence de certains motifs. En particulier, la présence de 13-mers CCNCCNTNNCCNC, reconnus par la protéine PRDM9 peut être utilisée pour prédire la localisation des hotspots de recombinaison sur le génome humain (Baudat et al. 2010, Myers et al. 2008).

Ces hotspots se situent aux mêmes loci dans les différentes populations

humaines (Hinch et al. 2011) mais ne sont pas du tout conservés par rapport aux chimpanzés (Ptak et al. 2005, Winckler 2005), et n'ont été actifs que depuis relativement récemment (entre 800 000 et 400 000 ans, Lesecque et al. (2014)). De manière générale, seules les forces génomiques créent de la diversité génétique par augmentation du nombre de mutations et d'haplotypes. Une mutation qui apparaît dans une population peut ensuite suivre trois voies différentes : (i) disparaitre de la population, (ii) se répandre dans la population jusqu'à fixation, c'est-à-dire jusqu'à ce que le ou les autres allèles disparaissent, ou bien (iii) osciller en fréquence sans jamais ni disparaître, ni se fixer. Les autres forces vont modifier la fréquence des mutations dans les populations.

1.2 Les facteurs démographiques influençant la diversité génétique

1.2.1 La dérive génétique et la taille des populations

La dérive génétique désigne la variation stochastique des fréquences alléliques dans une population en l'absence de sélection naturelle (Nei 1987, Wright 1931). Dans ces conditions, et sous isolement génétique (absence de migration), l'ensemble des allèles présents dans une génération proviennent d'un échantillonnage aléatoire des allèles dans le réservoir constitué par la génération précédente, et la probabilité de transmission d'un allèle d'une génération à l'autre dépend de la taille efficace de la population N_e, c'est-à-dire du nombre d'individus pouvant se reproduire dans une population où les rencontres sont aléatoires. Ainsi, une mutation qui apparaît dans une population humaine et donc diploïde a une probabilité de fixation égale à $\frac{1}{2N_e}$, et l'espérance du temps de fixation de cette mutation est égal à $4N_e$ générations. La dérive génétique influe donc sur les fréquences alléliques et provoque une diminution de la diversité génétique.

Selon ce modèle, les changements de taille d'une population ou histoire démographique vont avoir un impact direct sur sa diversité génétique. Une augmentation de la taille de la population (expansion) entraîne une plus faible dérive génétique et une augmentation de la diversité génétique de la population, alors qu'une réduction de la taille de la population ou goulot d'étranglement provoque une forte dérive génétique, avec la disparition ou la fixation d'un certain nombre

d'allèles, et une diminution de la diversité génétique. L'effet fondateur, qui se produit lorsqu'un petit groupe d'individus se sépare d'une population mère pour fonder une nouvelle population, est un cas de forte dérive génétique qui provoque une diminution importante de la diversité génétique et une variation considérable des fréquences des allèles.

1.2.2 Isolement, migration, flux génique

L'isolement désigne l'absence d'interfécondité entre deux populations. Il peut se produire soit par isolement géographique (distance, présence de barrière géographique), soit par isolement reproductif (par exemple, pour des raisons culturelles, Richerson and Boyd (2005)). L'isolement crée de la différenciation génétique entre les populations, avec l'apparition de mutations spécifiques à l'une ou l'autre des populations (Mayr 1963). La migration correspond au déplacement d'un individu ou d'un groupe d'individus d'une population vers une autre. Contrairement à la dérive génétique, la migration n'a pas d'impact sur la fréquence allélique au niveau de l'espèce, mais entraîne une augmentation de la diversité génétique de la population receveuse si les migrants participent à la diversité génétique de la génération suivante. On parle alors de flux génique (Cavalli-Sforza 1966, Cavalli-Sforza and Bodmer 1971, Cavalli-Sforza and Feldman 2003, Kimura and Weiss 1964, Wright 1943).

1.2.3 L'histoire démographique des populations humaines

Nous venons de le voir, les modifications de la taille d'une population et les événements de migration influencent sa diversité génétique. Il est aujourd'hui possible de faire le chemin inverse, et de se servir de l'étude la variabilité du génome humain, ainsi que des données archéologiques pour établir l'histoire démographique et migratoire des populations humaines. L'analyse phylogéographique du génome mitochondrial et du chromosome Y ont d'abord permis d'identifier l'origine africaine de l'espèce humaine (Cann et al. 1987, Cavalli-Sforza and Feldman 2003, Ingman et al. 2000, Thomson et al. 2000). Les datations génétiques et les enregistrements fossiles s'accordent à dater l'apparition de notre espèce il y a environ 200 000 ans en Afrique (Chen et al. 1995, Ingman et al. 2000, McDougall et al. 2005, Santos-Lopes et al. 2007). D'après l'étude des plus vieux fossiles retrouvés hors d'Afrique (Mellars 2006) et les datations génétiques à partir de données mitochondriales (Macaulay 2005,

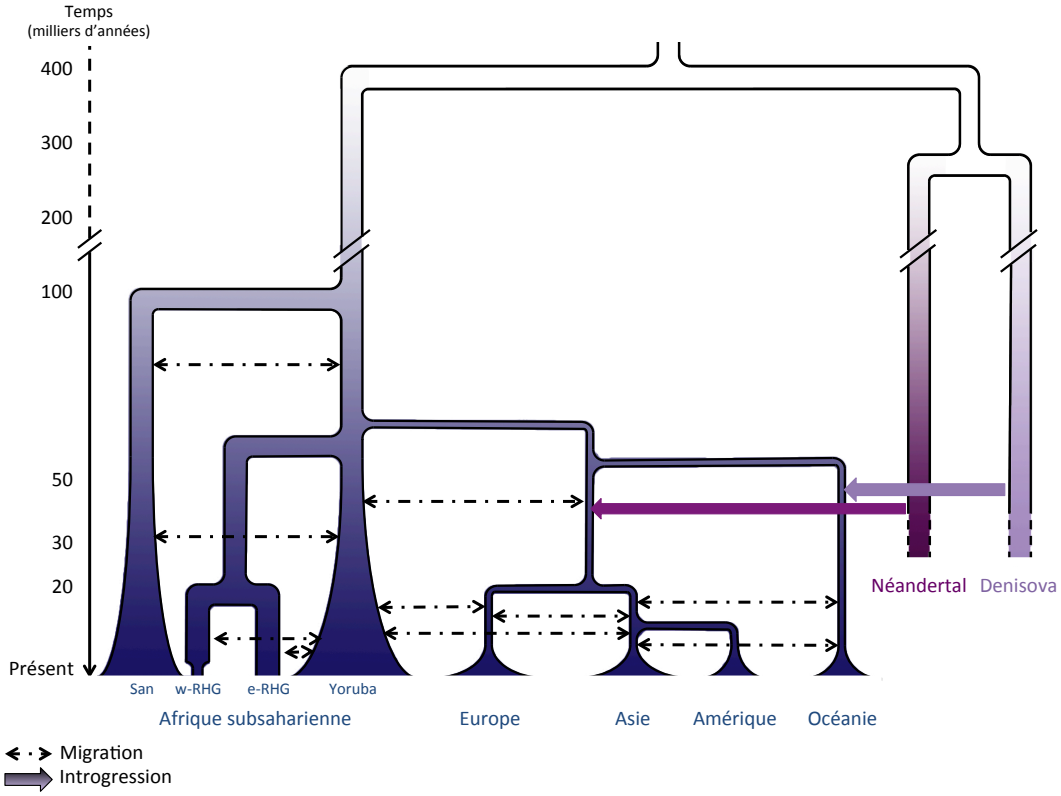


Fig. 2 Histoire démographique des populations humaines.
Les branches en bleu désignent la lignée humaine, les branches en violet, celles des Hommes archaïques Neandertal et Denisova. Les changements de taille efficace des populations sont donnés à titre indicatif (expansion modérée ou forte, goulot d'étranglement). Les variations d'épaisseur des branches ne représentent les facteurs réels d'expansions et de réductions. Pour l'Afrique sub-saharienne, des exemples de populations représentatives de chaque régime démographique sont notés en-dessous de chaque branche. w-RHG : chasseurs-cueilleurs Pygmées de l'Ouest de l'Afrique Centrale. e-RHG : chasseurs-cueilleurs Pygmées de l'Est de l'Afrique Centrale.

Quintana-Murci et al. 1999) et autosomales (Excoffier et al. 2013, Fagundes et al. 2007, Gravel et al. 2011, Hellenthal et al. 2008, Laval et al. 2010, Schaffner et al. 2005), de petits groupes de populations seraient ensuite sortis d'Afrique il y a entre 50 000 et 75 000 ans par la péninsule arabique avant de coloniser toute la planète (Océanie, Eurasie, Amérique). La plus grande diversité génétique des populations africaines et le fait que la diversité génétique des populations non africaines soit une sous-partie de la diversité génétique africaine corroborent ce modèle (Excoffier 2002, The 1000 Genomes Project 2010, The International HapMap 3 Consortium 2010, The International HapMap Consortium 2007).

Les génomes d'Hommes archaïques, obtenus grâce à l'avancée des techniques de séquençage de l'ADN ancien ont fourni des preuves que des événements d'admixture se sont produits entre Néandertal et les ancêtres des populations eurasiennes il y a

environ 50 000 à 60 000 ans (Fu et al. 2014, Green et al. 2010, Prüfer et al. 2014). Une cartographie des fragments de génome de Néandertal présents dans le génome des Hommes modernes a révélé qu'ils représentent entre 1,5 à 2,1% du génome des Européens et des Asiatiques, et que Néandertal a plus contribué au génome des populations asiatiques (Sankararaman et al. 2012, Vernot and Akey 2014, Wall et al. 2013), révélant un processus d'admixture complexe (Vernot and Akey 2015). De plus, 2 à 8% du génome des populations asiatiques et océaniennes d'Homme moderne proviendraient de Denisova (Prüfer et al. 2014, Reich et al. 2010, 2011). Enfin, le génome des populations africaines porte des traces d'admixture avec une population d'Hommes archaïques non identifiée (Hammer et al. 2011, Lachance et al. 2012, Plagnol and Wall 2006).

Ces éléments, ainsi que le croisement de données de génotypage et de séquençage avec des résultats de simulations, ont permis d'établir un modèle de variation des tailles des populations humaines résumant la diversité génétique humaine (figure 2). On distingue ainsi une séparation entre populations africaines et non africaines il y a 50 000 à 75 000 ans. Les populations africaines ont ensuite traversé un épisode d'expansion modérée, alors que la sortie d'Afrique s'est accompagnée d'un ou plusieurs goulots d'étranglement. Enfin, les populations asiatiques et européennes se sont séparées il y environ 20 000 ans, avant de connaître une expansion forte (Excoffier et al. 2013, Fagundes et al. 2007, Gravel et al. 2011, Laval et al. 2010, Schaffner et al. 2005). Des méthodes récentes ont permis de cartographier à une échelle plus fine les nombreux événements de migration et d'admixture ayant rythmé l'histoire démographique humaine (Alexander et al. 2009, Schiffels and Durbin 2014). Plus généralement, les populations humaines suivent donc un modèle d'isolement avec migration, où deux groupes d'individus se séparent par isolement géographique ou reproductif, puis échangent des migrants à un taux variable.

Chapitre 2

La sélection naturelle, action de l'environnement sur la diversité génétique

En plus des forces génomiques et démographiques que nous venons de voir, l'environnement peut aussi avoir un effet sur la diversité génétique (figure 1). Le mécanisme par lequel l'environnement influe sur la diversité génétique d'une population ou d'une espèce est appelé sélection naturelle. Ce concept a été défini pour la première fois par C. Darwin dans son livre le plus connu, *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. La sélection naturelle y est décrite comme le processus selon lequel un individu qui porte un trait avantageux dans un environnement donné aura un plus grand succès reproductif, et par conséquent, ce trait augmentera en fréquence au fil des générations dans la population. Cette théorie ainsi formulée repose sur trois grands principes : l'existence, dans une population, d'un trait variable selon les individus, l'héritabilité de ce trait, et la corrélation entre variation du trait et succès reproductif des individus. La sélection naturelle guide donc l'évolution des espèces et favorise leur survie et leur reproduction en permettant leur adaptation à l'environnement.

Cependant, la théorie darwinienne de l'évolution ne propose pas de mécanisme expliquant la transmission des traits au cours des générations. Il faudra attendre le début du XXe siècle et la redécouverte des travaux de G. Mendel sur l'hérédité pour que naisse la théorie synthétique de l'évolution, qui propose le gène comme unité de l'information génétique, et qui stipule que la sélection naturelle agit sur des mutations apparues par hasard au sein de la population. Les travaux de nombreux scientifiques au cours du XXème siècle, dont S. Wright, J.B.S Haldane, R.A. Fisher, G.H. Hardy,

W. Weindberg, T. Dobzhansky, W. D. Hamilton, R. Lewontin, E. Mayr et M. Kimura ont permis la mise en place d'un modèle général constituant une théorie ontologique et synthétique de l'évolution gouvernée par deux grands principes : d'une part, le hasard de l'évolution des fréquences des mutations dans une population (dérive génétique), et d'autre part, l'interaction entre génome et pressions environnementales (sélection naturelle).

2.1 Les différents types de sélection naturelle

Il existe trois régimes sélectifs principaux : la sélection positive, la sélection négative et la sélection balancée. Chaque régime a des conséquences différentes et spécifiques sur la diversité génétique locale qui constituent des signatures moléculaires (figure 3, Bamshad and Wooding (2003), Lohmueller et al. (2011), Nielsen (2005), Vitti et al. (2013)).

2.1.1 La sélection positive

Lorsqu'une mutation apparue par hasard confère un avantage sélectif à l'individu porteur, elle va augmenter en fréquence dans la population sous l'effet de la sélection positive plus rapidement que ne le ferait une mutation neutre sous l'effet du hasard. L'efficacité de la sélection dans une population, notée γ, est exprimée par la formule $\gamma = 2N_e s$, où s désigne le coefficient de sélection, qui correspond à l'avantage relatif procuré par l'allèle sélectionné par rapport à l'autre allèle, et N_e la taille efficace de la population. Le temps que la mutation va mettre à se fixer dans la population va dépendre de l'efficacité de la sélection, mais également du mode de transmission de l'allèle sélectionné (dominant, co-dominant ou récessif). Ainsi, une mutation dominante avec un coefficient de sélection élevé dans une population de grande taille se fixera plus rapidement qu'une mutation récessive avec un faible coefficient de sélection dans une petite population. La sélection positive entraîne des signatures moléculaires locales spécifiques, dont la diminution de la diversité génétique, l'augmentation du nombre d'allèles rares et très fréquents, l'augmentation du déséquilibre de liaison, et l'augmentation de la différenciation entre les populations (figure 3B, Lohmueller et al. (2011), Nielsen (2005), Pritchard et al. (2010), Scheinfeldt and Tishkoff (2013)).

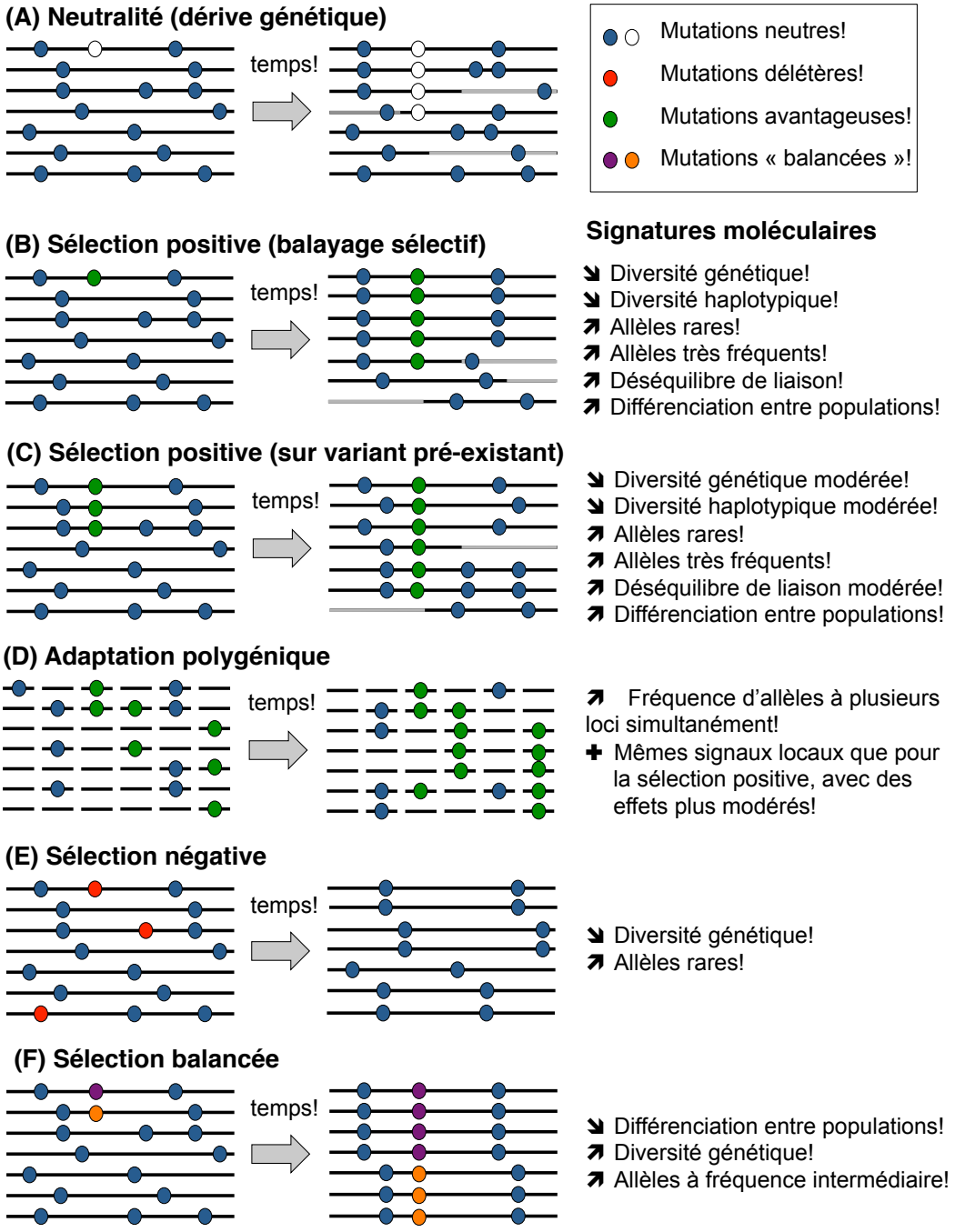


Fig. 3 Les différents régimes de sélection et leurs signatures moléculaires. Evolution d'une région génomique (**A**) sous neutralité, (**B**) sous sélection positive de type balayage sélectif, (**C**) sous sélection positive sur allèle pré-existant, (**D**) sous adaptation polygénique (**E**) sous sélection négative, (**F**) sous sélection balancée. Les points représentent les mutations. Les points bleus ou blancs sont des mutations évoluant sous neutralité, et les points d'autres couleurs des mutations évoluant sous sélection. Les traits représentent les haplotypes, et leur changement de couleur (de noir vers gris ou inversement), les événements de recombinaison.

On distingue plusieurs sous-types de sélection positive qui laissent des signatures moléculaires légèrement différentes sur le génome (figure 3C et D, Pritchard et al.

(2010), Scheinfeldt and Tishkoff (2013), Wollstein and Stephan (2015)). Dans le modèle classique, la sélection positive cible une mutation dès son apparition (*selective sweep*, figure 3B). On parle alors de balayage sélectif, et il peut être complet si l'allèle est fixé dans les populations (*classical sweep*), ou incomplet si l'augmentation en fréquence de l'allèle est toujours en cours (*ongoing sweep*). L'un des exemples le plus connu de balayage sélectif chez l'homme est celui de la persistance de la lactase, une enzyme qui permet la digestion du lactose et donc du lait. En effet, chez l'homme, la production de cette enzyme est maximale dans les semaines qui suivent la naissance, avant de diminuer progressivement, entraînant une incapacité à digérer le lait à l'âge adulte. Cependant, dans certaines populations d'Afrique, d'Europe et du Moyen-Orient, on observe une persistance de la production de la lactase à l'âge adulte chez plus de 80% des individus (figure 4, Itan et al. (2010)). Il s'agit d'un cas de sélection convergente, plusieurs mutations de la région régulatrice du gène *LCT* qui permettent la persistance de la production de lactase à l'âge adulte ayant en effet été la cible de la sélection positive dans différentes populations (Bersaglieri et al. 2004, Enattah et al. 2002, Tishkoff et al. 2007).

Il arrive cependant qu'une nouvelle mutation commence à évoluer en fréquence dans la population sous neutralité ou sous sélection négative faible avant de devenir avantageuse suite à un changement environnemental. On parle alors de sélection sur variant pré-existant (*selection on standing variation*, figure 3C). Enfin, l'avantage sélectif peut être conféré non pas par une seule mutation, mais par un ensemble de mutations à plusieurs loci (Novembre and Di Rienzo 2009, Pritchard and Di Rienzo 2010). Dans ce cas, la sélection positive va provoquer l'augmentation de la fréquence de plusieurs mutations à différents loci, et on parle de sélection polygénique (*polygenic adaptation*, figure 3D). Plusieurs phénotypes humains complexes sont sous sélection polygénique, comme la taille ou la résistance aux pathogènes. Ainsi, la taille d'un individu est un phénotype multi-génique, et plusieurs SNP relativement communs sont impliqués dans sa variabilité entre individus (Yang et al. 2010). Or un certain nombre d'allèles favorisant une plus grande taille présentent des fréquences plus élevées dans les populations du nord que dans celles du sud de l'Europe, indiquant l'existence d'une sélection faible mais ciblant plusieurs mutations favorisant une grande stature dans les populations nordiques (Turchin et al. 2012). Enfin, les signaux de sélection positive récente sont enrichis en mutations touchant des gènes impliqués

dans l'immunité et la réponse aux pathogènes (Daub et al. 2013), indiquant que ce trait à probablement également été la cible de sélection polygénique.

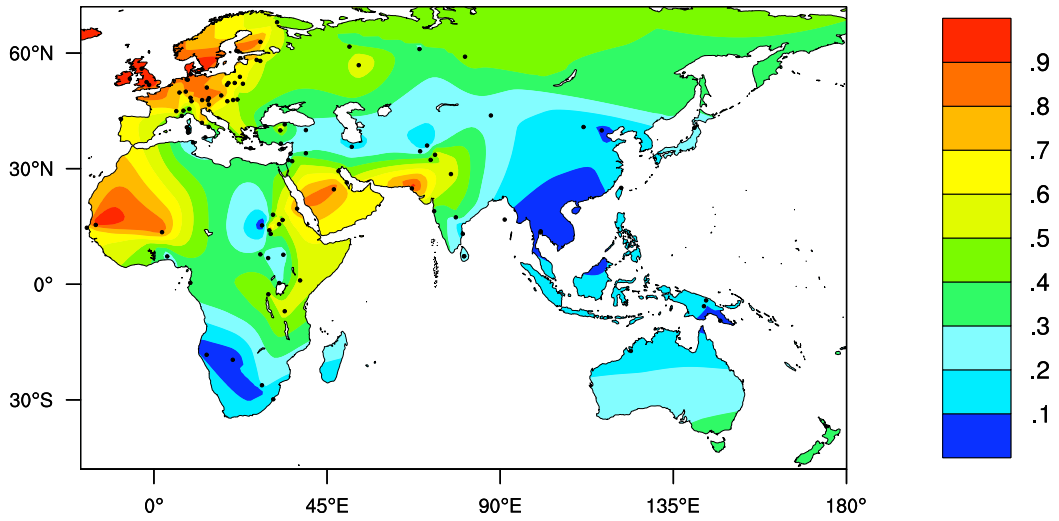


Fig. 4 Répartition du phénotype de persistance de la lactase à l'âge adulte. Carte tirée de Itan et al. (2010). Les couleurs représentent les fréquences du phénotype de persistance de la lactase estimés par interpolation de surface. Les points représentent les lieux d'échantillonnage.

2.1.2 La sélection négative

Contrairement à ce qui se passe lors de la sélection positive, lorsqu'une mutation apparue par hasard confère un désavantage pour la survie ou la reproduction à l'individu porteur, elle va diminuer en fréquence sous l'effet de la sélection négative jusqu'à disparaître. Le temps que la mutation va mettre à disparaître de la population dépend des mêmes paramètres que pour la sélection positive : le coefficient de sélection, qui correspond cette fois-ci au désavantage relatif procuré par l'allèle sous sélection négative par rapport à l'autre allèle, la taille efficace de la population, et le mode de transmission de l'allèle sélectionné. Ainsi, un allèle récessif mettra plus de temps à disparaître d'une population qu'un allèle dominant, puisqu'il ne devient délétère qu'à partir du moment où il est présent en deux copies chez un individu. De même, un allèle délétère peut atteindre une fréquence intermédiaire dans une population de petite taille du fait de la faible efficacité de la sélection négative et de la forte dérive génétique. Le goulot d'étranglement suivi par une expansion forte et récente des populations non africaines expliquerait l'excès de mutations fortement délétères chez les populations européennes par rapport aux populations africaines

(Casals et al. 2013, Lohmueller et al. 2008, Tennessen et al. 2012). Le bottleneck en lui-même provoque une diminution de l'efficacité de la sélection, permettant l'augmentation en fréquence de mutations délétères. Cet effet est ensuite amplifié par l'expansion rapide des populations non africaines, à l'origine du « *surfing* », processus au cours duquel les mutations délétères s'accumulent sur le front de l'expansion (Edmonds et al. 2004, Klopfstein et al. 2006, Peischl et al. 2013, Travis et al. 2007). Cependant, il existe à l'heure actuelle un débat sur le fait que les modifications récentes de la taille des populations dans l'histoire humaine aient eu un impact sur l'efficacité de la sélection naturelle (Casals et al. 2013, Fu and Akey 2013, Lohmueller 2014, Tennessen et al. 2012), certaines études n'ayant pas trouvé de preuves en ce sens (Do et al. 2015, Simons et al. 2014). Les signatures moléculaires de la sélection négative comprennent la diminution locale de la diversité génétique et l'augmentation du nombre d'allèles rares (figure 3E, Lohmueller et al. (2011), Nielsen (2005)).

On distingue deux types de sélection négative. Si la sélection négative est forte, la mutation très délétère sera purgée rapidement de la population. On parle alors de sélection purificatrice. Dans les cas de sélection plus mesurée ou sélection négative faible, une mutation moins délétère peut se maintenir à basse fréquence dans la population. Si la sélection négative se produit de manière récurrente sur plusieurs SNP dans la même région, elle provoque une diminution de la diversité génétique locale par disparition des haplotypes liés aux mutations délétères. C'est la sélection « d'arrière-plan » (*background selection*, Charlesworth (2012), Cutter and Payseur (2013), Nielsen (2005)). La sélection négative est le régime le plus fréquent de sélection dans le génome humain, et cible la majorité des gènes (Bustamante et al. 2005), mais également des régions non géniques (Asthana et al. 2007). Les éléments non codants sous contrainte semblent avoir un taux de renouvellement plus élevé que les régions codantes, suggérant une sélection négative plus récente (Rands et al. 2014, Ward and Kellis 2012a). Plus généralement, des études récentes s'accordent pour évaluer à environ 7 à 9% la proportion du génome humain sous contrainte, dont un peu plus de la moitié est conservée au sein des mammifères (Davydov et al. 2010, Rands et al. 2014, Ward and Kellis 2012a).

Parmi les gènes sous sélection négative, ceux impliqués dans les fonctions liées au cytosquelette et à la matrice extra-cellulaire semblent avoir été particulièrement contraints (Bustamante et al. 2005). On trouve parmi les gènes liés à l'immunité un

certain nombre d'exemples de fortes contraintes évolutives. Ainsi, la majorité des protéines de la famille des NLR, des NALP, et un sous-ensemble des interférons de type I et II sont sous forte sélection purificatrice (résumés dans Quintana-Murci and Clark (2013)). En particulier, le seul gène de la famille des interférons de type II, *IFNG*, qui code l'interféron γ impliqué dans la défense de l'hôte contre les infections ne présente aucune mutation non synonyme dans les populations étudiées (Manry et al. 2011). Cela suggère que toute mutation entraînant un changement de la séquence codante de ce gène jouant un rôle clef dans l'immunité est fortement délétère. Il faut également souligner l'exemple de la famille des *Toll-like receptors* (*TLR*, récepteurs de type Toll), ayant un rôle de senseur des pathogènes. Les membres intra-cellulaires de cette famille, spécialisés dans la reconnaissance des ADN et ARN de virus et de bactéries (*TLR3*, *TLR7*, *TLR8*, *TLR9*) sont sous sélection purificatrice (Barreiro et al. 2009), alors que les contraintes pesant sur les *TLR* membranaires sont beaucoup plus relâchées.

2.1.3 La sélection balancée

La sélection balancée agit sur plusieurs allèles à un même site, en favorisant la co-existance de plusieurs allèles. Ce régime de sélection provoque localement une diminution de la différenciation entre populations, et, contrairement à la sélection positive et négative, une augmentation de la diversité génétique et du nombre d'allèles à fréquences intermédiaires (figure 3F, (Charlesworth 2006, Nielsen 2005)). Il existe plusieurs cas de sélection balancée.

Dans le premier cas, l'état hétérozygote est plus avantageux que les deux états homozygotes. On parle alors de superdominance. Un exemple bien connu de superdominance est celui de l'allèle de l'hémoglobine S (HbS) en Afrique. Cet allèle à l'état homozygote est responsable de la drépanocytose, une maladie mendélienne qui affecte le transport de l'oxygène dans le sang et provoque des anémies. Malgré ses conséquences très délétères, cet allèle est présent à forte fréquence dans certaines régions où le paludisme est endémique (Hedrick and Thomson 1983). En effet, à l'état hétérozygote, cet allèle permet une meilleure résistance à l'infection par *Plasmodium falciparum*, parasite responsable du paludisme, sans avoir beaucoup d'effet sur le transport d'oxygène (Allison 1954, 1961).

Dans le deuxième cas, l'avantage octroyé par un phénotype dépend de sa fréquence

relative par rapport aux autres phénotypes. On parle alors de sélection dépendante de la fréquence, et elle peut être positive ou négative, selon que l'avantage procuré par le phénotype augmente ou diminue avec sa fréquence dans la population. Le système HLA (*human leucocyte antigène* ou antigènes des leucocytes humains) est une famille de gènes qui permettent de reconnaitre les marqueurs du "soi" et de présenter les antigènes provenant de pathogènes ou de cellules modifiées (par exemple cancéreuses) aux lymphocytes T pour déclencher la réponse immunitaire (Katz and Benacerraf 1976). Elle présente de nombreuses mutations et évolue sous sélection dépendante de la fréquence négative, c'est-à-dire que c'est la rareté de l'allèle qui donne le caractère avantageux (Raymond et al. 2005, Traherne et al. 2006).

Enfin, dans le troisième cas, la sélection balancée peut résulter d'une oscillation du génotype le plus avantageux dans le temps (au cours de la vie de l'individu) ou dans l'espace (en fonction de l'environnement). C'est la sélection variant dans le temps ou l'espace. Le gène *ABO*, qui code les groupes sanguins et dont les fréquences des trois allèles A, B et O varient beaucoup selon les populations humaines, évolue sous l'un ou l'autre de ces modèles potentiellement dans le cadre d'une co-évolution avec les pathogènes présents dans l'intestin (Ségurel et al. 2013).

La sélection balancée peut provoquer le maintien du polymorphisme allélique pendant une période très longue. Ce type de sélection balancée « à long terme » (*long-term balancing selection*) est caractérisé par la présence d'allèles trans-espèces (allèles présents dans plusieurs espèces) bien plus longtemps après leur divergence qu'attendu sous neutralité (Key et al. 2014). C'est ainsi le cas du système HLA, dont certains allèles sont apparus il y a plus de 40 millions d'années (Klein et al. 1993) et des allèles A et B du gènes *ABO*, apparus il y a environ 20 millions d'années chez les primates (Segurel et al. 2012). Il existe aussi des événements de sélection balancée plus récents comme le cas de HbS spécifiques de l'Homme.

La sélection naturelle peut donc affecter la diversité génétique de l'espèce humaine de plusieurs manières, mais contrairement aux forces génomiques et démographiques, elle a un effet localisé à la région génomique sous sélection. Dans cette thèse, je vais me concentrer sur l'étude des effets de la sélection positive sur le génome.

2.2 Retracer l'effet de l'environnement sur la diversité génétique : les marques laissées par la sélection positive sur le génome

L'ensemble des tests statistiques permettant de retracer l'histoire évolutive d'une région génomique ou d'une mutation en génétique des populations se basent sur trois postulats : un taux de mutation constant, aussi connu sous le nom d'horloge moléculaire, une taille de population constante et l'équilibre entre mutation et dérive, le nombre d'allèles perdus par dérive génétique étant compensé par le nombre d'allèles créés par mutation. En effet, l'observation d'une mutation isolée n'apporte aucune information sur son histoire évolutive. Il faut la considérer dans son contexte génomique. L'hypothèse nulle stipule donc que le génome évolue sous neutralité, et que les régions sous sélection présentent des signatures moléculaires s'écartant de la neutralité. Pour détecter ces régions, on cherche donc celles pour lesquelles l'hypothèse nulle est rejetée. Dans la suite de ce chapitre, je vais me concentrer sur les tests permettant de détecter la sélection positive. Il en existe deux grands types, les tests inter-spécifiques et les tests intra-spécifiques, tous basés sur l'étude des SNP, et permettant de détecter des événements de sélection plus ou moins récents en fonction des signatures moléculaires étudiées (Nielsen 2005, Nielsen et al. 2007, Sabeti et al. 2006, Vitti et al. 2013).

2.2.1 L'apport des comparaisons inter-spécifiques

Pour détecter la sélection positive dans notre espèce, on peut comparer certaines régions du génome humain à leurs homologues chez d'autres hominidés. Ces données inter-spécifiques permettent, en appliquant le principe de parcimonie, d'identifier pour un SNP donné l'allèle ancestral et l'allèle dérivé et d'obtenir une évaluation qualitative de l'âge de la mutation. Il existe trois principaux tests inter-spécifiques permettant de détecter des marques de sélection positive dans le génome exploitant l'accumulation de différences fixées entre deux espèces (divergences, par exemple entre l'homme et le chimpanzé), et du nombre de polymorphismes au sein de l'espèce humaine, résultant d'événements de sélection positive très anciens, ayant eu lieu dans la lignée humaine (figure 5, Sabeti et al. (2006)).

Parmi eux, deux sont basés sur l'étude des régions codantes et comparent

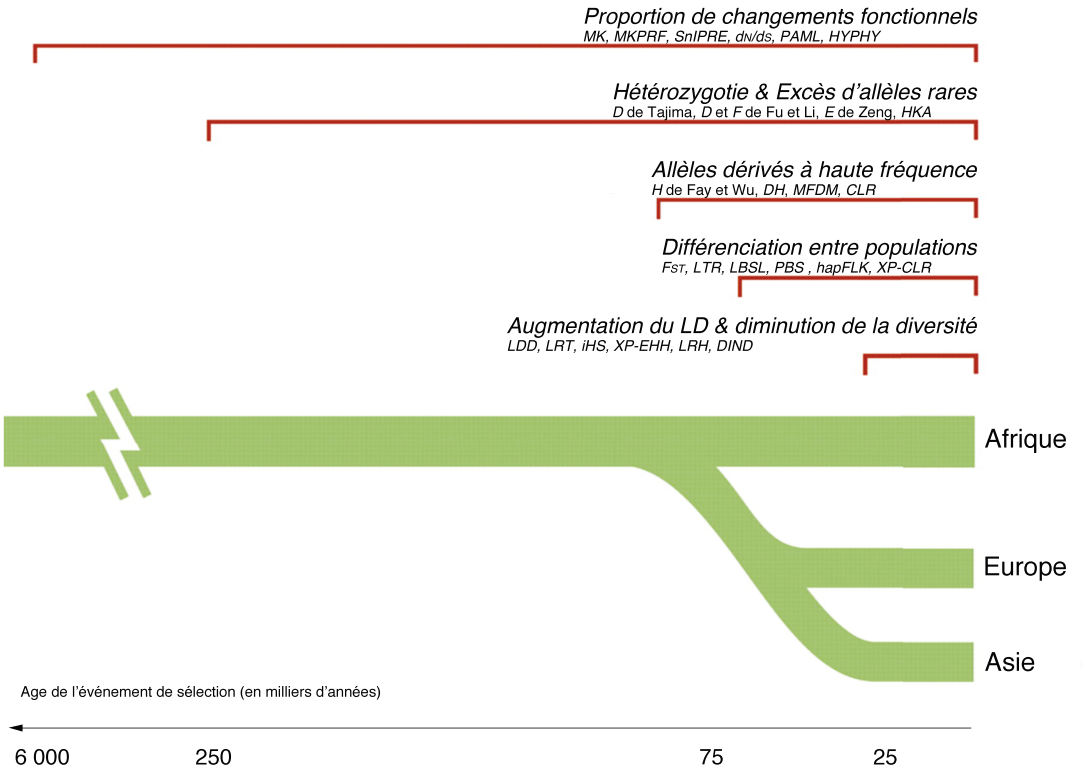


Fig. 5 Détecter la sélection positive : échelles de temps.
Adapté de Sabeti et al. (2006). Les intervalles en rouge représentent l'échelle de temps sur laquelle chaque groupe de statistique peut détecter des événements de sélection positive

le nombre de mutations fonctionnelles (non synonymes) et non fonctionnelles (synonymes). Le ratio $\frac{dN}{dS}$ (Yang and Bielawski 2000) et ses dérivés prenant en compte des informations phylogénétiques comme PAML (Yang 2007) et HYPHY (Pond et al. 2005) comparent le nombre de divergences synonymes et non synonymes. Le test MK (McDonald and Kreitman 1991) et ses deux variants MKPRF (Sawyer and Hartl 1992) et SnIPRE (Eilertson et al. 2012), quant à eux, comparent le nombre total de divergences et de polymorphismes synonymes (ne changeant pas la séquence d'acides aminés de la protéine) et non synonymes. Enfin, un troisième type de test, HKA (Hudson et al. 1987), compare directement le nombre de divergences au nombre de polymorphismes. Dans tous les cas, si un excès de polymorphismes et de divergences à un locus donné du génome par rapport à d'autres régions peut signer un événement de sélection positive, il peut aussi indiquer une pseudogénéisation ou l'action de la sélection dite « d'arrière plan », provoquée par la présence de mutations légèrement délétères à faible fréquence (tableau 1, Eyre-Walker and Keightley (2009), Messer and Petrov (2013a)).

2.2.2 L'étude du spectre de fréquence allélique

Sous les hypothèses de la théorie neutraliste de l'évolution, les fréquences alléliques au sein d'un échantillon d'individus suivent une distribution donnée (Tajima 1983). L'un des effets de la sélection positive est la modification du spectre de fréquences alléliques, avec notamment un enrichissement en mutations à très faibles ou très fortes fréquences (figure 3B). Il existe plusieurs tests permettant de déterminer si le spectre de fréquences alléliques observé correspond à celui attendu sous neutralité : le D de Tajima (Tajima 1989), le D et le F de Fu et Li (Fu and Li 1993), le H de Fay et Wu (Fay and Wu 2000) et le E de Zeng. Ces tests sont tous basés sur des comparaisons des estimateurs de θ, le taux de mutation de la population. Théoriquement, $\theta = 4N_e\mu$, où μ est le taux de mutation par génération. Ce taux ne peut être calculé directement à partir de données de polymorphisme d'un échantillon, mais il peut être estimé en utilisant différentes méthodes, qui donnent toutes des valeurs comparables si l'hypothèse de neutralité est respectée.

Watterson a le premier proposé un estimateur de θ, noté $\hat{\theta}_W$. Si S représente le nombre de sites polymorphes dans la population et n la taille de l'échantillon (nombre de chromosomes), alors :

$$\hat{\theta}_W = \frac{S}{a_n} \text{ avec } a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

Tajima a montré en 1989 qu'on pouvait estimer θ à partir du nombre de différences entre chaque paires de chromosomes. Si on note d_{ij} le nombre de différences entre le chromosome i et le chromosome j, alors :

$$\hat{\theta}_\pi = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}}{\binom{n}{2}}$$

En 1993, Fu et Li démontrent que le nombre de mutations dérivées présentes en un seul exemplaire dans l'échantillon ou singletons notée η_e est un estimateur de θ.

En 1995, Fay et Wu proposent un autre estimateur noté $\hat{\theta}_H$. Si ξ_i désigne le nombre de mutations pour lesquelles l'allèle dérivé est présent à i exemplaires dans l'échantillon, alors :

$$\hat{\theta}_H = \frac{\sum_{i=1}^{n-1} i^2 \xi_i}{\binom{n}{2}}$$

Enfin, en 2006, Zeng construit un nouvel estimateur noté $\hat{\theta}_L$. Alors :

$$\hat{\theta}_L = \frac{1}{n-1} \sum_{i=1}^{n-1} i\xi_i$$

Par construction, chacun de ces estimateurs est sensible à des modifications d'une partie différente du spectre de fréquence allélique. Ainsi, $\hat{\theta}_W$ et η_e sont sensibles aux variations de la proportion d'allèles rares, alors que $\hat{\theta}_\pi$ est sensible aux variations de la proportion d'allèles à fréquences intermédiaires, et $\hat{\theta}_H$ et $\hat{\theta}_L$ aux variations de la proportion d'allèles très fréquents.

Le D de Tajima, le D et le F de Fu et Li, le H de Fay et Wu et le E de Zeng comparent différents estimateurs. Ils sont égaux à 0 si le spectre de fréquence allélique de la région séquencée est celui attendu sous neutralité, et un écart à 0 de l'une ou l'autre de ces statistiques renseigne sur une déviation dans différentes parties du spectre par rapport à la neutralité.

Ainsi, le D de Tajima détecte les variations de la proportion d'allèles à fréquences intermédiaires. S'ils sont moins fréquents qu'attendu, D_{Tajima} sera négatif et inversement.

$$D_{Tajima} = \frac{\hat{\theta}_\pi - \hat{\theta}_W}{\sqrt{Var(\hat{\theta}_\pi) - Var(\hat{\theta}_W)}}$$

Le D et le F de Fu et Li vont permettre de détecter des variations de la proportion d'allèles rares. S'ils sont plus fréquents qu'attendu, D_{FuetLi} et F_{FuetLi} seront négatifs et inversement.

$$D_{FuetLi} = \frac{\hat{\theta}_W - \eta_e}{\sqrt{Var(\hat{\theta}_W) - Var(\eta_e)}}$$

$$F_{FuetLi} = \frac{\hat{\theta}_\pi - \eta_e}{\sqrt{Var(\hat{\theta}_\pi) - Var(\eta_e)}}$$

Enfin, le H de Fay et Wu et le le E de Zeng permettent de détecter des variations de la proportion d'allèles à haute fréquence. S'ils sont plus fréquents qu'attendu, $H_{FayetWu}$ et E_{Zeng} seront négatif et inversement.

$$H_{FayetWu} = \hat{\theta}_\pi - \hat{\theta}_H$$

$$E_{Zeng} = \frac{\hat{\theta}_L - \hat{\theta}_W}{\sqrt{Var(\hat{\theta}_L) - Var(\hat{\theta}_W)}}$$

Sous sélection positive, on observe un défaut de mutation à fréquences intermédiaires, et une augmentation des mutations rares et à haute fréquence (figure 3B). D_{Tajima}, D_{FuetLi}, F_{FuetLi} et E_{Zeng} seront donc négatifs. Cependant ces résultats peuvent aussi refléter l'existence d'événements de sélection purificatrice ou d'expansions de populations (tableau 1). Un moyen de faire la différence entre sélection négative et positive est donc d'utiliser $H_{FayetWu}$, qui sera également négatif sous sélection positive, comme dans le test *DH* qui combine les deux (Zeng et al. 2006). Tous ces tests permettent de détecter des événements de sélection positive anciens, ayant eu lieu dans l'espèce humaine ou avant (figure 5, Sabeti et al. (2006)).

2.2.3 La différenciation entre populations

Sous sélection positive locale, une mutation augmente en fréquence rapidement dans la population où elle est avantageuse mais continue d'évoluer sous neutralité dans les autres populations, provoquant une augmentation de la différenciation entre population à cette mutation. La statistique F_{ST} permet de mesurer la différenciation entre populations (Wright 1943, 1965). Sous neutralité, le F_{ST} est déterminé par la dérive génétique (Cavalli-Sforza 1966, Excoffier et al. 1992, Hudson et al. 1992, Lewontin and Krakauer 1973, Weir and Cockerham 1984). Dans le cas d'un événement de sélection positive locale, le F_{ST} de la mutation sélectionnée augmente (Bamshad and Wooding 2003, Barreiro et al. 2008, Cavalli-Sforza 1966). L'un des tests classiques utilisant cette propriété est celui de Lewontin-Krakauer (*LKT*, (Lewontin and Krakauer 1973)). Une autre statistique, *hapFLK*, est basée sur le même principe, mais calcule la différenciation entre populations des haplotypes (Fariello et al. 2013), et est insensible à la structuration des populations. Un fort F_{ST} peut donc refléter un événement de sélection positive relativement récent dans l'une des deux populations (figure 5, Sabeti et al. (2006)), mais aussi une différenciation entre les deux populations par isolement (tableau 1), et sa distribution globale dans le génome dépend fortement de l'éloignement géographique des populations comparées (Hutchison and Templeton 1999).

L'avantage du F_{ST} est qu'elle permet de détecter directement le variant sous sélection. Cependant elle ne permet pas de déterminer dans quelle population la sélection positive a agi. Pour cela, il existe de nouveaux tests comme *PBS* (*population branch statistics*, Yi et al. (2010), Zhang et al. (2005)), *LSBL* (*locus-specific branch*

lengths, Shriver et al. (2004)) et *LRT* (*likelihood ratio test*, Bhatia et al. (2011)). Ils utilisent le F_{ST} pour générer des arbres de distance génétique entre au moins deux populations à une mutation donnée, permettant ensuite l'identification de la population dans laquelle l'événement de sélection a eu lieu par comparaison de la longueur des branches.

2.2.4 Les variations locales de la longueur des haplotypes

Lorsqu'une mutation apparaît dans un haplotype donné, elle est en déséquilibre de liaison avec les autres allèles présents sur le même chromosome. Au cours des générations, la recombinaison va peu à peu casser ce déséquilibre de liaison et augmenter la diversité locale associée à cet allèle. Sous neutralité, une mutation augmente lentement en fréquence dans la population. Une mutation à haute fréquence présente donc une forte diversité locale et un faible déséquilibre de liaison avec les mutations voisines. Au contraire, une mutation sous sélection positive augmente rapidement en fréquence et va donc présenter à fréquence équivalente une plus faible diversité locale et un déséquilibre de liaison avec les mutations voisines plus fort (figure 3). Il existe aujourd'hui plusieurs statistiques capables de détecter des événements de sélection positive en se basant sur ces deux signatures.

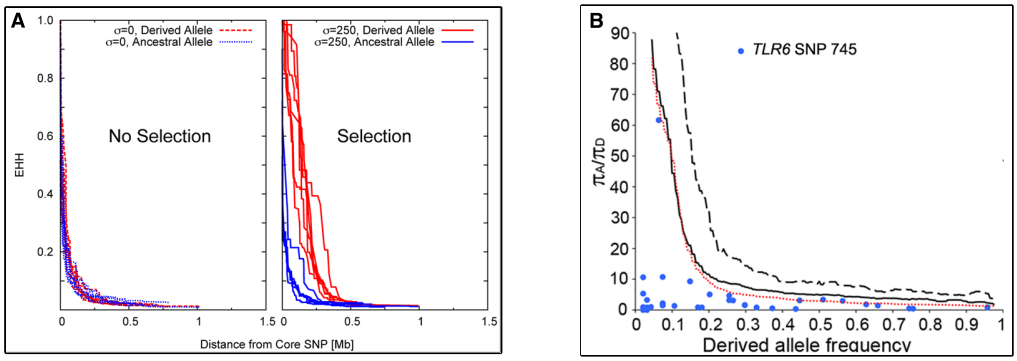


Fig. 6 Principe des statistiques basées sur la longueur des haplotypes.
(**A**) Effet de la sélection positive ciblant l'allèle dérivé sur le profil de diminution d'*EHH*. Tiré de Voight et al. (2006). Le premier panel présente les courbes d'*EHH* observées sous neutralité, et le second les courbes observées sous sélection positive de l'allèle dérivé (en rouge). (**B**) Principe de *DIND*. Tiré de Barreiro et al. (2009). Sous neutralité, *DIND* suit une distribution donnée : la ligne pleine noire et la ligne rouge donnent le $95^{ème}$ et le $99^{ème}$ centile selon le modèle démographique de Voight et al. (2006), la ligne pointillée noire donne le $95^{ème}$ centile selon un modèle démographique considérant une taille constante de la population. Le SNP 745 du gène *TLR6* présente une signature de sélection positive.

Les statistiques détectant l'augmentation locale du déséquilibre de liaison associée

à l'allèle sous sélection positive, hormis le LDD (*Linkage-desequilibrium decay*, Wang et al. (2006)) sont toutes basées sur le calcul de l'EHH (*extended haplotype homozygosity*, Sabeti et al. (2002)), qui mesure la diminution de l'homozygotie moyenne entre haplotypes en se déplaçant latéralement depuis la mutation d'intérêt. Elles permettent de détecter uniquement des événements de sélection positive très récents (<30 000 ans, figure 5, Sabeti et al. (2006)).

Pour un site situé à la position d_0 ayant deux allèles, l'allèle ancestral a_A et l'allèle d'intérêt a_D, l'EHH pour l'allèle d'intérêt à une distance x de ce site noté $EHH(x, a_D)$ correspond à la probabilité que deux chromosomes choisis par hasard dans un échantillon présentent le même haplotype sur l'ensemble de la région allant de d_0 à $d_0 + x$, et est donc compris entre 0 et 1. L'EHH d'un allèle dérivé sous sélection positive diminuera plus lentement que celui de l'allèle ancestral correspondant (figure 6).

Le test *LRH* (*long range haplotype*, Sabeti et al. (2002)) est basé sur le ratio de EHH calculés pour l'haplotype d'intérêt d'une part et pour l'ensemble des autres haplotypes d'autre part. Il consiste à comparer la valeur du ratio pour la mutation d'intérêt et à celle d'autres régions choisies au hasard et supposées évoluant sous neutralité. Une mutation sous sélection présentera un ratio significativement plus élevé. Sa généralisation, $WGLRH$ (Zhang et al. 2006), utilise une distribution de ce ratio sur l'ensemble du génome comme distribution nulle.

La statistique *iHS* (*integrated haplotype score*, Voight et al. (2006)) est aussi dérivé de l'EHH. Pour un site situé à la position d_0, soit L_d et L_g les distances à droite et à gauche pour lesquelles EHH est inférieur à 0,05 pour les allèles ancestraux et dérivés. On peut alors intégrer EHH :

$$iHH_D = \int_{d_0 - L_g}^{d_0 + L_d} EHH(x, a_D) \mathrm{d}x$$

$$iHH_A = \int_{d_0 - L_g}^{d_0 + L_d} EHH(x, a_A) \mathrm{d}x$$

et calculer l'iHS selon la formule ci-dessous :

$$iHS = ln\left(\frac{iHH_A}{iHH_D}\right)$$

Sous neutralité et après normalisation, iHS suit une loi normale centrée réduite. Sous sélection positive, qui provoque une augmentation du déséquilibre de liaison associé

à l'allèle sélectionné uniquement, iHS s'écarte significativement de cette distribution. $XP-EHH$ est autre statistique qui fonctionne sur le même principe que iHS, mais permet de comparer le déséquilibre de liaison autour d'une mutation entre deux populations (Sabeti et al. 2007).

Une autre statistique, *DIND* (*derived intraallelic nucleotide diversity*, Barreiro et al. (2009)), est basée sur la comparaison de la diversité nucléotidique locale des haplotypes associés à chaque allèle. Elle a été créée spécialement pour utiliser des données de séquençage. Même si *DIND* utilise les données du spectre de fréquence allélique et non la longueur des haplotypes, je l'ai placée dans cette section à cause de l'homologie qu'elle présente avec les statistiques basées sur l'EHH en terme de logique (comparaison des haplotypes ancestraux et dérivés), ainsi que d'ancienneté des événements de sélection détectés. Soit n_A et n_D le nombre de chromosomes portant respectivement l'allèle ancestral et l'allèle dérivé, d_{ij} le nombre de différences entre deux chromosomes i et j portant l'allèle ancestral et d_{kl} le nombre de différences entre deux chromosomes k et l portant l'allèle dérivé. Les diversités haplotypiques associées à l'allèle ancestral π_A et à l'allèle dérivé π_D sont alors égales à :

$$\pi_A = \sum_{i=1}^{n_A-1} \sum_{j=i+1}^{n_A} d_{ij} \text{ et } \pi_D = \sum_{k=1}^{n_D-1} \sum_{l=k+1}^{n_D} d_{kl}$$

DIND est alors calculé suivant la formule $DIND = \frac{\pi_A}{\pi_D}$.

Sous neutralité, DIND suit une distribution nulle, que l'on peut obtenir soit par simulation, soit à partir du calcul de DIND sur les données de séquençage de régions du génome évoluant sous neutralité. A une mutation dérivée sous sélection positive, qui entraîne une diminution de la diversité haplotypique environnante, DIND est significativement plus élevé qu'à d'autres mutations ayant une fréquence allélique similaire (figure 6).

Il existe donc une grande diversité de statistiques permettant de détecter diverses signatures de la sélection positive à partir de différents types de données génomiques (tableau 1). Cependant, chacune de ces statistiques prises séparément ne détecte la sélection positive que sur une échelle de temps et pour une gamme de fréquences alléliques restreintes. Enfin, l'usage de certaines statistiques comme celles basées sur l'étude du spectre de fréquence allélique ou celles basées sur EHH, permet seulement d'identifier des régions sous sélection et non les mutations responsables du signal. Pour répondre à ces limitations, on voit apparaître depuis quelques années des tests

basés sur des combinaisons de statistiques tels le *CMS* (*composite of multiple signals*, Grossman et al. (2013, 2010)). Ils tirent profit de plusieurs signatures moléculaires de la sélection positive à la fois, élargissant ainsi la portion du spectre de fréquence allélique sur laquelle il est possible de détecter la sélection et permettant de gagner en précision spatiale.

2.2.5 Distinguer démographie et sélection positive

Je l'ai évoqué dans les paragraphes précédents, les statistiques basées sur des comparaisons inter-spécifiques, sur l'analyse du spectre de fréquences alléliques et sur la différenciation entre populations sont sensibles à l'histoire démographique des populations (tableau 1). Or les populations humaines ne sont pas de taille constante (figure 2), les Africains montrant une variabilité génétique compatible avec une expansion modérée, alors que les Européens et les Asiatiques ont une variabilité génétique compatible avec un ou plusieurs goulots d'étranglement suivis d'une expansion importante.

Pour distinguer les effets de l'histoire démographique et de la sélection naturelle, on peut tirer profit du fait que la première affecte l'ensemble du génome, alors que la seconde n'affecte que localement la diversité génétique autour de la mutation sélectionnée. Pour détecter la sélection positive dans les populations humaines, on peut donc comparer les valeurs obtenues sur une région candidate en utilisant une statistique donnée à une distribution nulle intégrant les effets de la démographie. Celle-ci peut être obtenue en calculant les valeurs de cette même statistique sur des régions du génome supposées évoluer sous neutralité, et qui reflètent donc uniquement l'histoire démographique de la population. Si l'histoire démographique de la population est connue, cette distribution peut également être obtenue par simulation d'un modèle démographique réaliste (Fagundes et al. 2007, Gravel et al. 2011, Laval et al. 2010, Schaffner et al. 2005, Voight et al. 2005).

Il existe également des statistiques basées sur l'étude du spectre de fréquence allélique et corrigeant directement en interne les effets de la démographie, comme le *MFDM* (Li 2011) qui étudie le déséquilibre de l'arbre des fréquences alléliques à un loci ou le *CLR* (*composite likelihood ratio*, Nielsen et al. (2005b)), et son équivalent comparant les fréquences alléliques dans plusieurs populations l'$XP-CLR$ (Chen et al. 2010) qui combinent les scores obtenus pour plusieurs sites dans une

région donnée. Enfin, la comparaison des valeurs des statistiques directement entre sites fonctionnels (sites codants ou régulateurs) et apparemment non fonctionnels permet également d'identifier des régions sous sélection positive (Barreiro et al. 2008, Bustamante et al. 2005, The 1000 Genomes Project 2010, The International HapMap Consortium 2007, Voight et al. 2006). Dans tous les cas, l'utilisation de ces tests nécessite une annotation précise et fiable du génome pour distinguer les régions potentiellement fonctionnelles (régulatrices ou codant des protéines) des régions évoluant sous neutralité.

Table 1 Résumé des principales statistiques utilisées pour détecter la sélection positive.

Signatures moléculaires	Données utilisées	Facteurs confondants	Tests
Excès de polymorphismes et divergences aux sites non-synonymes par rapport aux sites synonymes	Séquences de plusieurs loci chez l'humain et dans une espèce de Primate voisine	Sélection «d'arrière-plan», Pseudogénéisation	*MK*, *dN/dS*, PAML, HYPHY
Excès de polymorphismes par rapport aux divergences	Séquences de plusieurs loci chez l'humain et dans une espèce de Primate voisine	Sélection «d'arrière-plan», Pseudogénéisation	*HKA*
Ecart au spectre de fréquences alléliques dérivées attendu sous neutralité	Séquences de plusieurs loci chez l'humain et dans une espèce de Primate voisine	Goulot d'étranglement	*H* de Fay et Wu
Ecart au spectre de fréquences alléliques attendu sous neutralité	Séquences de plusieurs loci chez l'humain	Sélection négative, Expansion	*D* de Tajima, *D* et *F* de Fu et Li, CLR
Différenciation entre populations	Génotypage de plusieurs populations humaines	Isolement des populations	F_{ST}, *PBS*, *LSBL*, *LRT*
Augmentation du déséquilibre de liaison local associé à l'allèle ancestral	Séquence d'une région d'environ 1Mb ou d'un génome entier	Sélection «d'arrière-plan»	*EHH*, *iHS*, *LRH*, *WGLRH*
Diminution de la diversité locale associée à l'allèle sélectionné	Séquence d'une région d'environ 1Mb ou d'un génome entier		*DIND*
Combinaison de plusieurs signatures	Données de séquençage		*CMS*

Chapitre 3

La sélection positive à l'heure des études génomiques

Comme de nombreuses autres espèces, l'Homme a été confronté au cours de son histoire à des changements environnementaux (climats, régimes alimentaires, pathogènes entre autres) auxquels il a dû s'adapter, notamment lors de la colonisation du monde ou suite à l'invention de l'agriculture. Nous avons vu comment le hasard a forgé sa diversité génétique (cf. chapitre 1). Des travaux récents ont permi de détecter des traces de sélection positive à quelques loci du génome, et d'établir qu'un certain nombre de phénotypes ont été la cible de l'évolution adaptative récemment (Harris 2015, Nielsen et al. 2007, Sabeti et al. 2007, Vitti et al. 2013, Wollstein and Stephan 2015).

Avec l'essor du génotypage puis du séquençage à haut débit (The 1000 Genomes Project 2010, The 1000 Genomes Project Consortium 2012, The International HapMap 3 Consortium 2010, The International HapMap Consortium 2005, The International HapMap Consortium 2007) , les études cherchant à détecter la sélection positive à l'échelle du génome se sont multipliées, utilisant diverses classes de statistiques comme les comparaisons entre espèces (Clark et al. 2003, Nielsen et al. 2005a), les écarts au spectre de fréquence allélique (Akey et al. 2002, Carlson et al. 2005, Hinds et al. 2005, Jin et al. 2011, Kelley et al. 2006, Oleksyk et al. 2008, The International HapMap Consortium 2007, Weir et al. 2005), la différenciation entre populations (Barreiro et al. 2008, Williamson et al. 2007), la variation en longueur des haplotypes (Chen et al. 2010, 2015, Kimura et al. 2007, Pickrell et al. 2009, Sabeti et al. 2007, Tang et al. 2007, The International HapMap Consortium 2007, Voight et al. 2006, Wang et al. 2006) ou une combinaison de ces différentes méthodes (Grossman

et al. 2013). Basées dans leur immense majorité sur la détection des « *outliers* », régions présentant un profil aberrant par rapport à l'ensemble du génome, elles se sont le plus souvent attachées à détecter les événements de balayage sélectif, plus facilement détectables de par l'ampleur des signatures laissées par l'augmentation rapide d'un allèle venant d'apparaître dans une population.

3.1 Exemples de sélection positive dans les populations humaines

Parmi les signaux de sélection positive détectés, on distingue des événements d'adaptation causés par divers types de pressions. Si celles-ci ne sont pas toujours faciles à identifier, comme dans le cas des multiples événements de sélection ciblant des gènes impliqués dans la spermatogenèse (Nielsen et al. 2005a), on soupçonne la sélection sexuelle de jouer un rôle non négligeable dans la sélection de certains phénotypes morphologiques. Une mutation du gène *HMGA2* associée à la taille dans la population générale est ainsi sous sélection en Europe Ayub et al. (2014), Weedon et al. (2007). Certains événements de sélection ciblant des allèles ayant de multiples conséquences phénotypiques pourraient être dû à l'action simultanée de plusieurs pressions de sélection. La mutation non synonyme (V370A) du gène *EDAR* est ainsi associé à l'épaisseur des cheveux et à la forme des dents chez l'humain (Barreiro et al. 2008, Bryk et al. 2008, Fujimoto et al. 2008) est un exemple intéressant. En effet, une induction de cette mutation chez la souris a révélé qu'elle était également associée avec l'augmentation du nombre de glandes sudoripares (Kamberov et al. 2013). Ce phénotype ayant été retrouvé chez l'Homme, il est possible que sa sélection en Asie soit le résultat d'une adaptation à l'humidité doublée d'une sélection sexuelle (Kamberov et al. 2013). De manière générale, de nombreuses études indiquent que le climat, le régime alimentaire et les pathogènes font partie des trois pressions sélectives majeures s'étant exercées récemment sur le génome humain et ont eu des effets importants sur la diversité phénotypique des populations humaines (Akey 2009, Barreiro and Quintana-Murci 2010, Nielsen et al. 2007, Vallender 2004).

3.1.1 Adaptation au climat

Il existe un certain nombre d'exemples d'adaptation des populations humaines au climat(Hancock et al. 2011, Wollstein and Stephan 2015). La plus connue est l'adaptation à l'ensoleillement et à l'exposition aux UV. On trouve en effet parmi les gènes sous sélection en Europe et en Asie, ceux associés à une pigmentation plus claire de la peau, des cheveux et des yeux comme *SLC24A5*, *MATP*, *KITLG*, *TYR*, *HERC2*, *OCA2*, *TPCN2* et *ASIP*. Ils ont probablement été sélectionnés car ils confèrent un avantage dans le cadre d'une faible exposition aux UV (Izagirre et al. 2006, Sabeti et al. 2007, Sulem et al. 2008, Wilde et al. 2014), bien qu'on ne puisse pas exclure le rôle de la sélection sexuelle pour la couleur des yeux et des cheveux. Au contraire, en Afrique, le gène *MC1R*, impliqué dans la production de mélanine, est sous forte sélection purificatrice, probablement car la mélanine est indispensable à la protection contre les dommages provoqués par l'exposition aux rayons UV sur l'ADN et la lyse du folate (Harding et al. 2000, Jablonski and Chaplin 2000). Le travail de Hancock et al. (2011) a démontré l'existence de signatures d'adaptation à d'autres facteurs climatiques, dont la latitude et la température, le taux de précipitations et l'humidité. Il est cependant important de noter que ces paramètres climatiques sont très corrélés entre eux.

Il existe d'autres exemples d'adaptation au climat, et notamment à l'hypoxie causée par la haute altitude. Ainsi, les allèles de *EGLN1*, *EPAS1* et *PPARA* entraînant une diminution de la concentration en hémoglobine sont sous balayage sélectif dans des populations vivant en altitude au Tibet (Beall et al. 2010, Simonson et al. 2010, Yi et al. 2010). Des mutations dans quatre autres gènes (*CBARA1*, *VAV3*, *ARNT2* and *THRB*) ainsi qu'un SNP intergénique, rs10803083, associés au même phénotype, sont également sous sélection positive dans les populations vivant sur les hauts plateaux éthiopiens (Alkorta-Aranburu et al. 2012, Scheinfeldt et al. 2012). Enfin, deux autres gènes impliqués dans des voies métaboliques associées à la vasodilatation et à la détection de l'hypoxie, respectivement *NOS2A* et *PRKAA1* montrent des traces d'adaptation en Amérique du Sud (Bigham et al. 2010, 2009). Il s'agit d'un cas de sélection convergente (Foll et al. 2014, Hancock et al. 2011). L'accumulation de données génomiques issues de populations vivant dans des conditions climatiques particulières avec une forte résolution géographique va permettre de détecter les mutations fortement différenciées entre populations et dont la fréquence est corrélée à

des facteurs climatiques, et ainsi d'en apprendre plus sur les pressions exercées par le climat sur la diversité humaine (Coop et al. 2010, Novembre and Di Rienzo 2009).

3.1.2 Adaptation aux changements de régimes alimentaires

Le régime alimentaire a également joué un rôle important sur la diversité phénotypique humaine (Luca et al. 2010). Il semble notamment que l'homme soit adapté à un mode de vie où les sources de nourriture sont incertaines (Neel 1962). Plusieurs SNP associés au diabète de type 2, à l'obésité et à l'augmentation de l'index de masse corporelle sont sous sélection positive chez l'Homme, et particulièrement en Océanie (Klimentidis et al. 2011). Le risque de développer un diabète serait donc une conséquence secondaire de l'adaptation à la variation de la disponibilité des ressources alimentaires, maintenant que ces populations sont dans un environnement où la nourriture est abondante. Au contraire, une mutation dans le gène *TCF7L2*, associée à une diminution de l'incidence du diabète de type 2 et à une réduction de l'indice de masse corporelle est sous sélection positive en Europe et en Asie. La datation de l'événement d'adaptation suggère qu'il est contemporain du développement de l'agriculture en Europe et doit donc conférer un avantage dans cet environnement (Klimentidis et al. 2011).

L'invention de l'agriculture et son expansion à la quasi-totalité des populations humaines lors des derniers 10 000 ans semble par ailleurs avoir exercé de nombreuses pressions de sélection sur le génome des populations humaines. Les spécificités alimentaires liées au développement de types de cultures et d'élevages différents ont notamment été la source d'adaptations locales, notamment du métabolisme. L'exemple de la persistance de la lactase à l'âge adulte, dont l'expansion est concomitante de celle de l'agriculture au Moyen-orient, en Europe et en Afrique, a été abordé plus haut. Il existe également d'autres adaptations du métabolisme au changement de régime alimentaire. Ainsi, la mutation rs1229984 du gène *ADH1B* qui altère la capacité à digérer l'alcool et protège probablement contre l'alcoolisme (Dick and Foroud 2003), présente des signatures de balayage sélectif en Asie de l'Est et en Europe (Barreiro et al. 2008, Galinsky et al. 2015, Han et al. 2007), et son augmentation en fréquence dans les populations suit l'expansion de la culture du riz à l'est de l'Asie (Peng et al. 2010). De même, diverses mutations de *NAT2* causant une acétylation plus lente sont sous sélection positive dans les populations pratiquant

l'agriculture ou le pastoralisme, probablement en réponse à la diminution des folates dans le régime alimentaire (Luca et al. 2008, Patin et al. 2006).

Un certain nombre de mutations impliquées dans le goût et l'olfaction, deux sens particulièrement important dans l'alimentation, portent également des signatures de sélection positive. Par exemple, une mutation non synonyme dans *TAS2R16* provoquant une plus grande sensibilité à certains glycosides est sous sélection dans l'espèce humaine. Ce phénotype a probablement été sélectionné chez les ancêtres chasseurs-cueilleurs des populations humaines car il confère une protection contre les toxines cyanogènes présentes dans certaines plantes (Soranzo et al. 2005). Un certain nombre de récepteurs d'olfaction sont également sous sélection positive dans plusieurs populations humaines (Gilad and Lancet 2003, Gilad et al. 2003, Williamson et al. 2007). De façon générale, les gènes sous sélections sont enrichis en gènes impliqués dans le métabolisme des glucides, des graisses et de l'alcool, la perception du goût et l'olfaction (Barreiro et al. 2008, Voight et al. 2006), ce qui indique que de nombreux événements d'adaptation pourraient être liés au régime alimentaire.

3.1.3 Adaptation aux pathogènes

Enfin, l'exposition au pathogènes est probablement la pression de sélection la plus importante sur le génome humain (Barreiro and Quintana-Murci 2010, Casanova et al. 2013, Fumagalli et al. 2011, Quintana-Murci and Clark 2013, Siddle and Quintana-Murci 2014). En effet, avant l'apparition des vaccins et des antibiotiques, la moitié des enfants mouraient avant l'âge de 15 ans, et l'espérance de vie était d'environ 20 ans, expliquant pourquoi la sélection naturelle a particulièrement ciblé les gènes de défense contre les pathogènes. Nous avons vu de nombreux exemples de sélection négative ou balancée, mais environ 200 gènes liés à l'immunité montrent également des signatures de sélection positive (Barreiro and Quintana-Murci 2010). De nombreux pathogènes semblent avoir joué un rôle dans l'évolution des populations humaines : *Plasmodium falciparum*, le parasite responsable du paludisme, les bactéries responsables de la lèpre, de la tuberculose, du choléra, et de nombreux virus (Karlsson et al. 2014).

Des mutations dans des gènes impliqués dans la résistance au paludisme sont par exemple sous forte sélection positive dans les zones où le parasite est endémique (*FY*, aussi connu sous le nom de *DARC* et *CR1* Afrique sub-saharienne et *G6PD* en Asie du Sud-Est, Barreiro et al. (2008), Hamblin and Di Rienzo (2000), Louicharoen et al.

(2009), Sabeti et al. (2006), Tishkoff (2001)). La lèpre fréquente en Europe jusqu'à la fin du Moyen-Age a ensuite pratiquement disparue alors qu'elle reste présente en Inde et en Asie de l'Est (Barreiro et al. 2009, Boldsen 2009, Grossman et al. 2013). Cela pourrait s'expliquer par le développement d'une résistance génétique à la maladie dans la populations européenne, hypothèse confortée par la découverte d'événements de sélection positive récents ayant ciblé des allèles protégeant contre la maladie dans cette population, mais pas en Asie (Karlsson et al. 2014) De même, une mutation de *LARGE* qui permet de diminuer la sensibilité des cellules à l'infection par le virus de la fièvre de Lassa (Andersen et al. 2012, Sabeti et al. 2007) est sous sélection positive en Afrique de l'Ouest, et des gènes de la famille *TLR* tels que *TLR5* en Afrique et le cluster *TLR1-6-10* en Europe et en Asie (Barreiro et al. 2009, Grossman et al. 2013, Wlasiuk et al. 2009) ont été la cible de la sélection positive, probablement à cause de leur rôle de senseurs de bactéries à la surface des cellules, ce qui nécessite de s'adapter à de multiples pathogènes pouvant évoluer rapidement.

3.2 Apports et limites des études génomiques pour l'étude de la sélection positive

3.2.1 Intérêts des études « génome entier »

Les différentes études génomiques de la sélection positive ont permis la détection de centaines de régions sous évolution adaptative. Si on observe peu de chevauchement entre leurs différents résultats, probablement à cause de la multiplicité des statistiques employées et du fort taux de fausses découvertes (FDR, *False discovery rate*) inhérent à la méthode « *outliers* », ces études ont néanmoins permis de mieux comprendre l'histoire évolutive de l'espèce humaine (Akey 2009), notamment en montrant que la colonisation de la planète s'est accompagnée d'épisodes locaux d'adaptation de populations à des environnements particuliers. D'autres études ont permis de montrer que la sélection positive n'a pas eu le même impact dans toutes les populations. En effet, les populations africaines, qui ont une plus grande taille efficace que les populations européennes et asiatiques, présentent plus de mutations avantageuses. Au contraire, dans les populations non africaines, on observe plus de mutations avantageuses fixées, effet direct de la plus forte dérive génétique causée par la succession d'effets fondateurs à l'origine de la formation de ces populations (Coop

et al. 2009, Pickrell et al. 2009).

L'étude de la sélection positive à l'échelle du génome a également permis d'établir une liste de gènes potentiellement soumis à des pressions adaptatives (Li et al. 2014). Si certains de ces signaux de sélection désignent des mutations dont on connaît la fonction et le rôle adaptatif dans l'histoire évolutive des populations humaines, comme celles de la lactase ou des mutations protégeant du paludisme, beaucoup tombent dans des régions dont la fonction est inconnue, ou bien trop larges pour identifier précisément la mutation sous sélection (Grossman et al. 2010, Kelley and Swanson 2008). Or la détection de traces de sélection dans ces régions indique qu'elles abritent des mutations qui ont dû procurer et procurent peut-être toujours un avantage sélectif aux individus porteurs, et qui sont donc fonctionnelles, soit parce qu'elles modifient la séquence d'acide aminés et la fonction d'une protéine, soit parce qu'elles altèrent la régulation de l'expression de gènes. Pour cette raison, ces régions sont des candidats intéressants dans le cadre d'étude d'associations entre diversité génétique et diversité phénotypique.

Enfin, certaines études ont montré que les mutations associées à un risque de développer des maladies complexes, comme par exemple les maladies auto-immunes ou inflammatoires, sont enrichies en signaux de sélection positive (Barreiro and Quintana-Murci 2010, Blekhman et al. 2008, Nielsen et al. 2009, Raj et al. 2013, Ramos et al. 2014). La maladie cœliac en est un cas emblématique : de nombreux allèles augmentant le risque de développer la maladie sont sous sélection positive (Abadie et al. 2011, Sams and Hawks 2013). Il a été démontré qu'un de ces allèles, situé dans le gène *SH2B3*, sous sélection en Europe, entraîne une plus forte activation de la voie *NOD2* après stimulation de leucocytes par des lipopolysaccharides et permet donc probablement une meilleure réponse à l'infection par des bactéries. Si le mécanisme ayant conduit à la sélection de telles mutations est mal compris (est-ce que, par exemple, ces mutations ont été sélectionnées par le passé parce qu'elles conféraient une meilleure protection contre certains pathogènes ?), de telles observations sont intéressantes, en ce qu'elles indiquent que les régions sous sélections positives pourraient non seulement abriter des mutations fonctionnelles mais aussi des mutations candidates pour les études d'association avec les maladies complexes.

3.2.2 Le séquençage à haut débit, avantages et problèmes potentiels

Jusqu'à il y a cinq ans, les études de sélection positive se faisaient sur des données de génotypage génome entier tels que *Perlegen* (Hinds et al. 2005) ou *The International HapMap Project* (The International HapMap 3 Consortium 2010, The International HapMap Consortium 2005, The International HapMap Consortium 2007) donnant accès à respectivement environ 1,5 et plus de 3 millions de SNP. Ces jeux de données sont appauvris en SNP présents à faible fréquence dans les populations par rapport à des données de séquençage, ce qui entraîne une sur-représentation des SNP à fréquence intermédiaire. Ce biais qui trouve sa source dans la méthode de détermination des SNP affecte toute les statistiques basées sur les spectres de fréquences alléliques, le déséquilibre de liaison et le F_{ST} et ne peut être que partiellement corrigé (Clark 2005), en particulier lorsqu'il est important (Kelley et al. 2006). Les techniques de séquençage à haut débit ont permis l'apparition de données de séquençage de plusieurs génomes entiers individuels tels que *The 1 000 Genomes Project* (The 1000 Genomes Project 2010, The 1000 Genomes Project Consortium 2012) et *Complete Genomics* (Drmanac et al. 2010), dépourvues de ce biais de détermination, puisque leur puissance de détection des SNP rares atteint 99%, et ont permis d'accéder à respectivement 38 et 12 millions de SNP, soit jusqu'à dix fois plus que les données de génotypage.

Malgré tous les progrès concernant la connaissance de la diversité génétique des populations humaines, il existe une controverse sur l'ampleur de l'effet de la sélection positive sur cette diversité. Si plusieurs études ont obtenu des résultats suggérant que la sélection positive a exercé une force évolutive non négligeable sur le génome humain (Barreiro et al. 2008, Bustamante et al. 2005, Enard et al. 2014, Jin et al. 2011, Luisi et al. 2015, The 1000 Genomes Project 2010, The International HapMap Consortium 2007, Voight et al. 2006), d'autres études ont suggéré que les signaux détectés pouvaient provenir, au moins partiellement, de facteurs confondants tels que la diminution de la diversité locale provoquée par l'élimination d'allèles délétères par sélection négative (Coop et al. 2009, Hernandez et al. 2011, Pritchard et al. 2010). Des études récentes utilisant des données de génotypage ou de séquençage ont ainsi conclu qu'un mode particulier mais très étudié de sélection positive, le balayage sélectif, a très peu participé à l'évolution récente des populations humaines (Granka et al. 2012,

Hernandez et al. 2011). La question de l'importance de la sélection positive et du balayage sélectif dans l'évolution humaine reste donc ouverte.

Chapitre 4

Les acteurs épigénétiques, sources de variabilité phénotypique

Nous venons de voir comment la diversité phénotypique trouve sa source dans la diversité génétique et peut être influencée par l'environnement. Pendant la seconde moitié du XXème siècle, de nombreuses recherches se sont concentrées sur l'étude des liens entre variations génétiques et phénotypiques, espérant pouvoir ainsi expliquer l'intégralité de la diversité observée entre différents individus. Or un certain nombre d'études épidémiologiques et l'arrivée des données de génotypage et de séquençage à haut-débit depuis la fin des années 90 ont mis en lumière l'impossibilité d'expliquer la variance de certains traits phénotypiques uniquement par des facteurs génétiques. Ainsi, les travaux réalisés sur des cohortes des jumeaux monozygotes ont montré que deux individus possédant un génome identique peuvent présenter des phénotypes différents, en particulier en terme de maladies complexes (Ballestar 2010, Petronis 2006, Tysk et al. 1988, Zdravkovic et al. 2002). De plus, des études épidémiologiques à grande échelle indiquent que le fait de développer un cancer dépend plus ou moins de facteurs génétiques en fonction du type de cancer (Dolinoy et al. 2007, Lichtenstein et al. 2000). Enfin, dans les deux cas, on note une influence importante de l'environnement (Boomsma et al. 2002, Chakravarti and Little 2003, Dolinoy et al. 2007, Willett 2002). Ces observations ont permis de remettre sur le devant de la scène la notion d'un mécanisme externe au génome, sensible à l'environnement et capable d'influencer les phénotypes sans modifier l'information génétique (Morange 2002) : l'épigénétique.

Le terme épigénétique, formé par ajout du préfixe grec "épi-" signifiant "au-dessus" à "génétique" a connu plusieurs définitions au cours de sa courte histoire,

qui ont toutes en commun de désigner l'étude des modifications phénotypiques non explicables par des mutations génétiques. Forgé par Waddington en 1942, il désigne au départ une branche de la biologie étudiant les mécanismes de causalité entre les gènes et leurs produits qui déterminent le phénotype d'un individu. Aujourd'hui, il désigne l'étude de « l'ensemble des changements d'activité des gènes qui sont transmis au fil des divisions cellulaires ou au fil des générations sans faire appel à des mutations de l'ADN » (Le Monde Science et Techno, 2012). Si le génome désigne l'ensemble du matériel d'une cellule portant l'information génétique (l'ADN chez l'homme), l'épigénome désigne l'état épigénétique de la cellule, c'est-à-dire l'état de l'ensemble des acteurs non génétiques participant à la régulation de l'expression des gènes. Les modifications épigénétiques sont cruciales au cours du développement en permettant, à partir d'une cellule-oeuf unique, de créer de nombreuses cellules dont le génome est identique, mais dont les épigénomes sont différents, permettant ainsi de créer des cellules présentant des caractéristiques et des fonctions différentes. Si l'épigénome varie d'un type cellulaire à l'autre, il peut aussi varier d'un individu à l'autre pour un même type cellulaire, permettant d'expliquer pourquoi, par exemple, alors que deux jumeaux monozygotes ont un génome identiques, il ne sont pas des copies exactes l'un de l'autre (Wong 2005). La diversité des profils épigénétiques pourrait donc expliquer une part importante des variations phénotypiques humaines.

4.1 Les différents acteurs épigénétiques

4.1.1 Les états chromatiniens et l'expression des gènes

Dans le noyau des cellules humaines, l'ADN est présent sous forme de chromatine, dont l'unité de base est un segment d'ADN enroulé autour d'un complexe protéique constitué de huit histones (Kornberg 1974). A cette configuration de base peuvent se rajouter d'autres protéines et des ARN (acides ribonucléiques), en particulier dans les cas de où les chromosomes sont extrêmement compactés, lors de la mitose, de la méiose ou dans les spermatozoïdes par exemple. Le degré de condensation de la chromatine est fortement relié au niveau d'expression des gènes.

On distingue deux types de chromatine. L'euchromatine ou « vraie chromatine » correspond à la chromatine non condensée et est composée uniquement d'ADN et d'histones. Peu structurée, elle permet aux protéines telles que les facteurs

de transcription et les ARN polymérases d'accéder à l'ADN et comprend les régions activement transcrites de l'ADN. L'hétérochromatine, quant à elle, désigne la chromatine condensée. L'ADN est peu accessible aux protéines et peu ou pas transcrit, et comprend deux sous-types : l'hétérochromatine constitutive (contenant un peu moins de 10% de l'ADN nucléaire (International Human Genome Sequencing Consortium 2004)), très condensée, présente dans tous les types cellulaires et contenant par exemple les centromères et les télomères, et l'hétérochromatine facultative, qui contient les gènes non exprimés et dont la composition en séquences d'ADN dépend des types cellulaires. Pendant l'interphase, la chromatine est organisée en domaines présentant différents niveaux de condensation et séparés les uns des autres par des isolateurs caractérisés par la présence du facteur de transcription CTCF (Ho et al. 2014, Labrador and Corces 2002, Phillips and Corces 2009). Ces états peuvent se transmettre tout au long des divisions cellulaires (Felsenfeld and Groudine 2003, Khorasanizadeh 2004).

La chromatine, au sein d'une cellule en interphase, est une structure dynamique et qui peut être remodelée notamment lors de la transcription (déplacement / replacement des histones lors du passage de l'ARN polymérase) ou afin de répression de l'expression de gènes (Wolffe and Guschin 2000). L'analyse de la structure de la chromatine via la recherche des sites hyper-sensibles à la DNase I, présents dans des régions où l'ADN n'est plus enroulé autour d'histones, a permis d'établir des cartes de la chromatine ouverte à l'échelle du génome. Ces études ont montré que l'état de la chromatine à un locus dépend du type cellulaire (Boyle et al. 2008, ENCODE Project Consortium 2012). De manière générale, le degré de compaction de la chromatine est déterminée par un certain nombre d'acteurs épigénétiques (figure 7), qui permettent ainsi de contrôler de façon fine le niveau expression des gènes et donc d'agir sur le phénotype.

4.1.2 Les différents acteurs épigénétiques

L'état de la chromatine est déterminé par deux principaux marqueurs : la méthylation de l'ADN et les modifications des queues N-terminales des histones (figure 7 et tableau 2). La méthylation de l'ADN est une modification covalente de l'ADN qui ne change pas l'information génétique. Dans l'espèce humaine, elle se fait par ajout d'un groupement méthyle sur des cytosines qui deviennent alors

des 5-méthylcytosines (5-mC), le plus souvent dans le contexte d'un di-nucléotide cytosine-guanine (CpG, cf. figure 8). La propriété symétrique de ces di-nucléotides est utilisée pour assurer la transmission des profils de méthylation eu cours des divisions cellulaires.

Les modifications des histones peuvent être réparties en au moins huit types, ayant des rôles différents dans la régulation de la transcription, la réplication, la réparation et la condensation de l'ADN (Kouzarides 2007). Les plus étudiées sont l'acétylation des lysines, généralement associée à l'euchromatine et à une transcription active, et la méthylation des lysines ou des arginines, associée soit à l'activation de l'expression des gènes, soit à la répression de la transcription et à l'hétérochromatine selon la position des résidus modifiés et la région génomique touchée (promoteurs, régions codantes, séquences répétées). Il semble que les marques d'histones soient transmises au travers des divisions cellulaires, même si les modalités de cette héritabilité sont mal connues (Goldberg et al. 2007, Hansen et al. 2008). La co-localisation de la méthylation de l'ADN avec certains types de modifications d'histones paraît constituer un code épigénétique déterminant l'état de la chromatine (Ho et al. 2014, Roadmap Epigenomics Consortium et al. 2015, Siegfried and Simon 2010). De par sa grande stabilité, sa facilité d'accès ainsi que son caractère informatif sur la régulation de la transcription (Burger et al. 2013, Reik 2007), la méthylation de l'ADN sera utilisée comme marqueur de l'état épigénétique et de la régulation des gènes dans cette thèse.

Ces dernières années, de plus en plus d'études ont mis en évidence le rôle d'autres acteurs épigénétiques, les ARN non codants (ARNnc) (Bernstein and Allis 2005) dans un certain nombre de processus via des interactions avec la chromatine (tableau 2) : la régulation de l'expression de certains gènes durant le développement et la différenciation cellulaire, l'empreinte parentale (Cao 2014) et le remodelage de la chromatine (Khalil et al. 2009). Xist, par exemple, est l'ARNnc le plus connu. Il est impliqué dans la répression aléatoire d'un des deux chromosomes X chez la femme à une étape du développement embryonnaire, répression qui ensuite transmise de façon stable à travers les divisions cellulaires (Brown et al. 1991). D'autres petits ARN non codants sont également impliqués dans la régulation de la transcription des gènes et la répression des éléments transposables. Parmi eux, les ARNpi (petits ARN interagissant avec les protéines Piwi) jouent un rôle dans la méthylation *de novo* de l'ADN chez les mammifères, fréquente lors de la gamétogenèse et des premiers

stades du développement embryonnaire (Aravin et al. 2008, Sigurdsson et al. 2012, Watanabe et al. 2011). Ils servent notamment à réprimer les éléments transposables, une étape cruciale au vu de l'effet particulièrement délétère que peut avoir l'insertion aléatoire d'un tel élément. Enfin, les microARN, petits ARN impliqués dans la régulation de l'expression des gènes, sont également souvent considérés comme des acteurs épigénétiques, bien que cette classification soit débattue car ils agissent en aval et non en amont de la transcription, par séquestration ou clivage des ARN messagers déjà synthétisés (Castel and Martienssen 2013). Dans tous les cas, les ARNnc impliqués dans les processus épigénétiques sont transmissibles à travers les générations cellulaires par répartition aléatoire dans les cellules filles lors de la division cellulaire (Bernstein and Allis 2005).

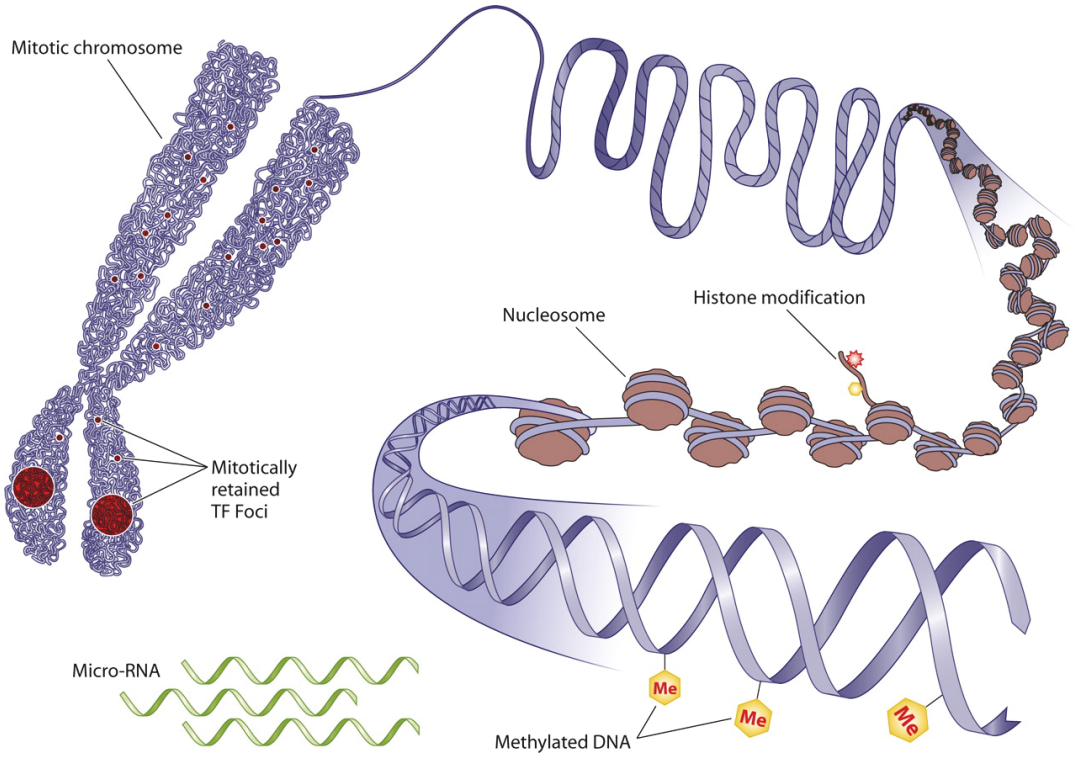


Fig. 7 Les différents acteurs épigénétiques. Figure tirée de Zaidi et al. (2010)

L'ensemble de ces acteurs interagissent pour réguler l'état de la chromatine et la transcription des gènes selon des modalités encore largement inconnues, sauf pour quelques cas particuliers. Un exemple intéressant d'interaction entre les différents acteurs épigénétiques est la répression des éléments transposables dans les cellules de la lignée germinale chez les mammifères. En effet, celle-ci nécessite la présence conjuguée de la marque d'histone H3K9me3 et de la méthylation de l'ADN, et

l'intervention d'ARNpi (Ha et al. 2014, Pezic et al. 2014, Sigurdsson et al. 2012). Une altération de la voie PIWI chez l'homme entraîne en effet la production d'ARNpi altérés et la diminution de la méthylation de certains éléments transposables dans les cellules de la lignée germinale et est associée à une spermatogenèse défectueuse et à des problèmes d'infertilité (Heyn et al. 2012). Plus généralement, le démarrage récent de programmes visant à caractériser l'état de différents marqueurs épigénétiques ainsi qu'à mesurer le niveau d'expression des gènes dans divers types cellulaires commence à apporter une vision plus générale des interactions au sein de l'épigénome (ENCODE Project Consortium 2012, Ernst et al. 2011, Ho et al. 2014, Roadmap Epigenomics Consortium et al. 2015) et de leur lien avec la régulation de la transcription (Yan et al. 2015), condition *sine qua non* pour comprendre les effets phénotypiques des variations épigénétiques.

4.2 La méthylation de l'ADN : genèse des profils et rôle

4.2.1 Les mécanismes de méthylation et de déméthylation chez l'Homme

Les cytosines méthylées se retrouvent dans deux types de configuration le long du génome humain : elles sont incluses soit dans un dinucléotide CpG, la méthylation est alors symétrique sur les deux brins et héritable à travers les divisions cellulaires (figure 8) ; soit dans des dinucléotides CpA et CpT (cytosine-adénosine et cytosine-thymine) et la méthylation alors asymétrique n'est pas transmissible au cours des divisions cellulaires et est diluée au cours du temps. Le premier scénario est largement majoritaire dans les cellules somatiques humaines mais la méthylation asymétrique est relativement fréquente dans les cellules souches embryonnaires (Ramsahoye et al. 2000). Des études récentes ont également montré son existence dans certains tissus somatiques (Pinney 2014). Il existe plusieurs mécanismes de méthylation de l'ADN impliquant des enzymes de la famille DNMT (ADN méthyltransférase). DNMT1 est responsable du maintien de la méthylation pendant les divisions cellulaires, où elle méthyle le brin nouvellement synthétisé par symétrie par rapport au brin parental, et joue aussi un rôle dans le maintien de la méthylation entre les divisions cellulaires

Table 2 Les acteurs épigénétiques : rôles et exemples.

Marques épigénétiques	Héritabilité au cours des divisions	Role épigénétique / Etat de la chromatine
Méthylation de l'ADN	Réplication au cours de la division par copie de l'état de méthylation du brin mère assurée par *DNMT1*	Empreinte parentale
		Hétérochromatine facultative, répression stable dans la lignée cellulaire
		Hétérochromatine constitutive, répression des éléments transposables, centromères, télomères
		Euchromatine, expression des gènes
		Euchromatine, repression des éléments transposables dans les introns et des promoteurs alternatifs ?
Modifications des queues d'histones	Mécanisme inconnu	Euchromatine, transcription active
		Hétérochromatine facultative, répression stable dans la lignée cellulaire
		Hétérochromatine constitutive, répression des éléments transposables, centromères, télomères
long ARNnc	Répartition dans les cellules filles au cours de la division	Régulation de l'expression de certains gènes pendant le développement
		Empreinte parentale
		Remodelage de la chromatine
ARNpi	Répartition dans les cellules filles au cours de la division	De novo methylation, répression des éléments transposables dans la lignée germinale et durant le développement
microARN	Répartition dans les cellules filles au cours de la division	Répression post-transcriptionelle de l'expression des gènes

(Schermelleh et al. 2007). DNMT3A et DNMT3B, en combinaison avec DNMT3L qui augmente l'affinité des deux premières enzymes pour l'ADN et leur activité, interviennent dans la méthylation *de novo* de l'ADN, et leur forte activité dans les cellules germinales et les cellules souches embryonnaires pourrait être responsable du fort taux de méthylation hors des sites CpG observé dans ces cellules (Okano et al. 1999).

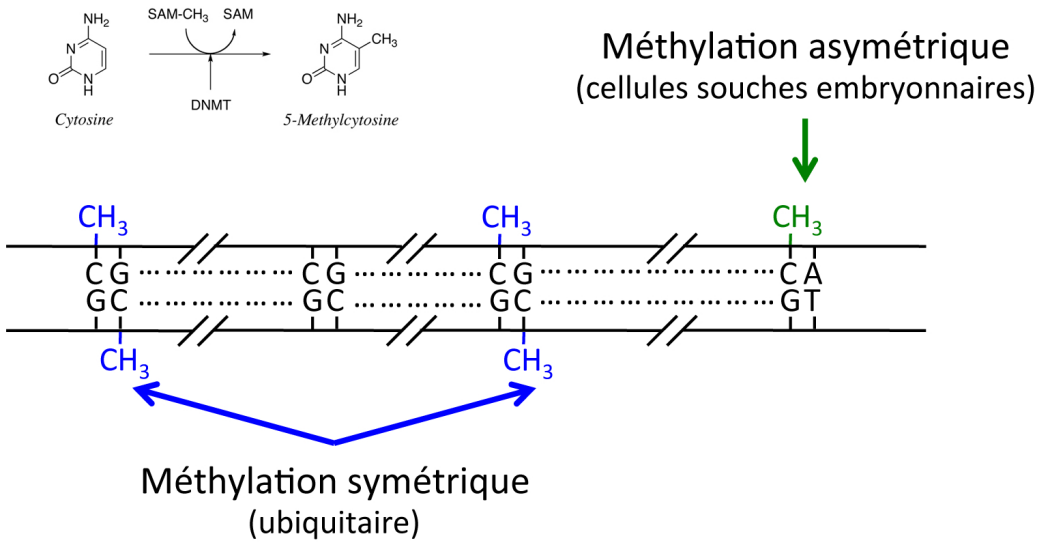


Fig. 8 La méthylation des cytosines. (**A**) Méthylation d'une cytosine par une DNMT. (**B**) Méthylation symétrique et asymétrique du génome humain.

Les mécanismes de déméthylation sont plus complexes et sont longtemps restés inconnus. Cependant la découverte de la famille de protéines TET (*ten-eleven translocation*, Tahiliani et al. (2009)) a permis de proposer deux types de méchanismes (figure refdemethylation). Après hydroxylation active des 5-mC en 5-hydroxyméthylcytosines (5-hmC), la déméthylation peut se produire soit par dilution passive au cours des divisions cellulaires, les 5-hmC étant beaucoup moins bien répliquées par DNMT1 que les 5-mC (Hashimoto et al. 2012, Valinluck and Sowers 2007), soit par réparation active. Cette deuxième voie implique des oxydations supplémentaires des 5-hmC en 5-formylcytosines (5-fC) et éventuellement en 5-carboxylcytosines (5-caC). Ces deux bases sont reconnues et excisées par l'enzyme TDG (*thymine DNA glycosylase*), provoquant la prise en charge du site par le système de réparation par excision de base, qui permet le rétablissement d'une cytosine non méthylée (He et al. 2011, Ito et al. 2011, Yu et al. 2012, Zhang et al. 2012). Ces machineries de méthylation et de déméthylation permettent une certaine souplesse du profil de méthylation notamment lors du développement embryonnaire.

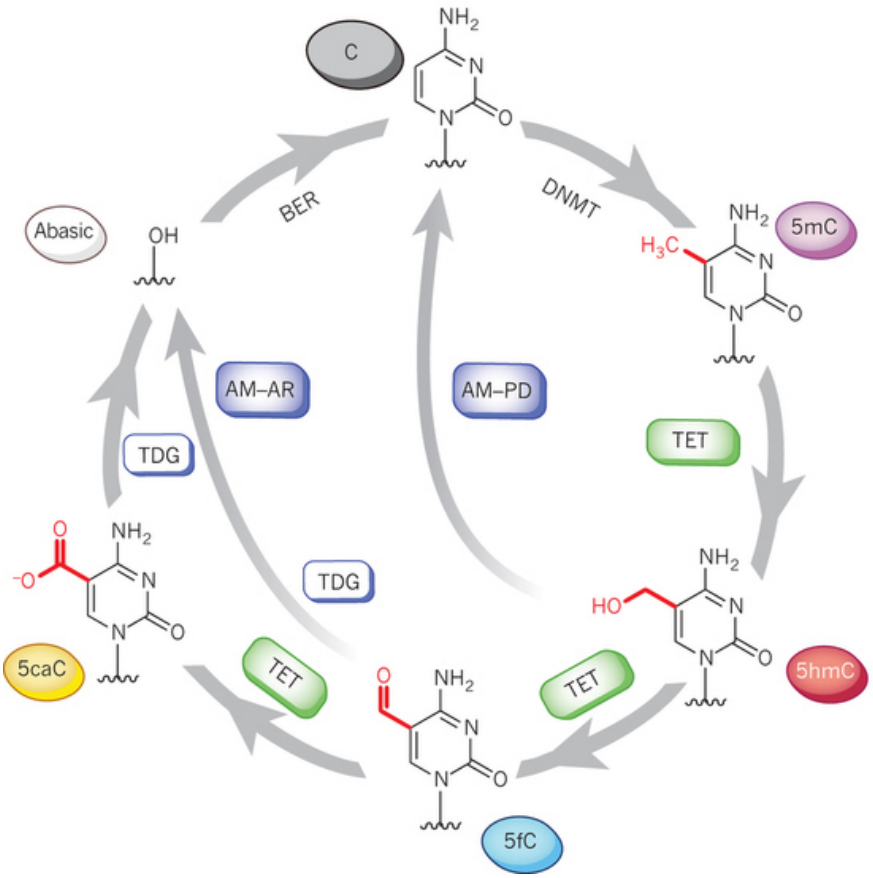


Fig. 9 Les mécanismes de déméthylation de l'ADN. Figure tirée de Kohli and Zhang (2013). On distingue deux voies après modification de la 5-mC en 5-hmC par l'enzyme TET : la déméthylation passive par dilution (AM-PD), et la déméthylation active par le système de réparation par excision de base (BER) impliquant l'enzyme TGD (AM-AR).

4.2.2 Le profil de méthylation de l'ADN chez l'humain : caractérisation et conservation

La méthylation de l'ADN n'est pas homogène le long du génome humain. Environ 1% des C et 70 à 80% des dinucléotides CpG sont méthylés dans les cellules somatiques (Ehrlich et al. 1982). Les 5-mC présentent un taux de mutation élevé, probablement parce que leur désamination les transforment non en uraciles comme les C, détectables et réparés par la machinerie cellulaire, mais en thymine (Scarano et al. 1967), provoquant un déficit en dinucléotides CpG dans le génome des vertébrés. On estime en effet qu'ils sont quatre à cinq fois moins fréquents qu'attendu chez l'humain (Jabbari and Bernardi 2004). Cependant, il existe des régions du génomes riches en dinucléotides CpG appelés les îlots CpG, généralement localisés dans les promoteurs des gènes, qui ont la particularité d'être très peu méthylés. Si on observe

une structuration à courte distance (environ 1 500 pb) des niveaux de méthylation, celle-ci est encore plus forte au niveau des îlots CpG (Bell et al. 2011). Le profil de méthylation des promoteurs dépend donc de la classe à laquelle ils appartiennent : ceux qui contiennent des îlots CpG (72%) sont généralement faiblement méthylés alors que les autres ont un niveau de méthylation variable (Saxonov et al. 2006). Le corps des gènes ainsi que les régions intergéniques, généralement pauvres en CpG, sont très méthylés (figure 10 A) et le sont d'autant plus que le gène est exprimé. Les éléments transposables, les télomères et les centromères présentent aussi un fort niveau de méthylation (Jones 2012). Enfin, il existe quelques régions hémi-méthylées (dont un des allèles est méthylé et l'autre non), dont les régions à empreintes parentales (Reik et al. 1987).

La méthylation de l'ADN est un mécanisme présent dans tous les règnes du vivant. En particulier, chez les eucaryotes, elle est médiée par des enzymes de la famille DNMT, qui est conservée entre les différentes espèces, même si elle est absente chez quelques-unes (Colot and Rossignol 1999). Cependant, les profils et le rôle de la méthylation de l'ADN varient grandement entre les protistes, les champignons, les plantes et les animaux. Chez les vertébrés, la méthylation concerne quasi uniquement des cytosines impliquées dans des dinucléotides CpG et le génome est majoritairement méthylé, à l'exception d'îlots CpG (Colot and Rossignol 1999, Jiang et al. 2014). S'il existe une conservation modérée des profils de méthylation entre les différentes espèces de vertébrés supérieurs (Jiang et al. 2014), plusieurs études ont montré que les mammifères présentent une très forte conservation des profils de méthylation. Ainsi, plus de 70% des locus orthologues et 90% des promoteurs des gènes homologues entre humains et souris ont un profil de méthylation fortement conservé (Eckhardt et al. 2006, Jiang et al. 2014). De même, seul un peu plus de 8% des sites de méthylation dans les promoteurs de gènes montrent des différences entre entre chimpanzés et humains (Pai et al. 2011). Enfin, une reconstruction du profil de méthylation de l'ADN de Néandertal et Denisova a trouvé une concordance de 99% entre les méthylomes d'un homme moderne et des deux hommes archaïques (Gokhman et al. 2014), confirmant la très grande conservation des profils de méthylation au cours de l'évolution.

4.2.3 Les rôles de la méthylation de l'ADN

La méthylation de l'ADN aux sites CpG intervient dans plusieurs processus liés à la répression de l'expression de certains éléments. Elle est notamment le médiateur principal de l'empreinte parentale en association avec d'autres acteurs épigénétiques (Li et al. 1993, Smallwood and Kelsey 2012). Ce terme désigne le phénomène de déséquilibre allélique de l'expression de certains gènes se traduisant par la transcription spécifique d'un seul des deux allèles parentaux lors de l'embryogenèse. Pour un gène, c'est toujours l'allèle de la même origine (paternelle ou maternelle) qui est exprimé. On répertorie aujourd'hui plus de 240 gènes montrant une expression mono-allélique couplée à une méthylation complète du promoteur de l'autre allèle (geneimprint, http ://www.geneimprint.com/site/home). Des études sur la souris montrent que l'empreinte parentale est nécessaire au développement normal de l'embryon, et suggèrent que ce mécanisme est responsable de la non viabilité des embryons obtenus par parthénogenèse chez les mammifères (Kawahara et al. 2007, Mann and Lovell-Badge 1984, McGrath and Solter 1984). La méthylation de l'ADN joue également un rôle important dans la répression de l'expression et le maintien dans l'hétérochromatine d'un certain nombre de régions comme les centromères, les télomères et les éléments transposables, ainsi que dans l'inactivation aléatoire d'un chromosome X chez les femmes (Smith and Meissner 2013). Ces inactivations se font via l'intervention d'une famille de protéines reconnaissant les 5-mCpG, les MBD (*methyl-CpG-binding domain*, Klose and Bird (2006)), qui recrutent des complexes répressifs constitués notamment d'enzymes permettant de méthyler le résidu H3K9, associé à l'hétérochromatine constitutive. De manière générale, la méthylation de l'ADN est nécessaire à la différenciation des cellules souches (Smith and Meissner 2013) et au développement normal de l'embryon, l'absence de DNMT fonctionnelles menant à une mortalité embryonnaire (DNMT1, Li et al. (1992)), peu après la naissance (DNMT3A) ou à un développement anormal de l'embryon menant à une mortalité in utero (DNMT3B, Okano et al. (1999)).

Il existe une corrélation complexe entre expression des gènes et niveau de méthylation du promoteur (Schübeler 2015). Si on observe globalement une corrélation négative entre les deux (Jones 2012), cette relation dépend d'abord de la classe à laquelle le promoteur appartient (Saxonov et al. 2006). En effet, les promoteurs contenant des îlots CpG sont généralement non méthylés, quel que soit le

niveau d'expression du gène, et sont particulièrement présents en amont des gènes de ménages qui sont exprimés de façon ubiquitaire. On observe cependant parfois une méthylation de tels promoteurs au cours du développement et de la différenciation cellulaire. Elle est alors associée avec la répression stable de l'expression du gène et son passage dans l'hétérochromatine (Jones 2012, Schübeler 2015). Les liens entre méthylation des promoteurs pauvres en CpG, qui sont particulièrement présents dans les gènes exprimés dans des types cellulaires spécifiques, et expression sont moins connus, mais la fixation de facteurs de transcription sur de telles régions entraîne une perte de la méthylation (Schübeler 2015). Plus généralement, les promoteurs des gènes très exprimés sont totalement dépourvus de méthylation (Bell et al. 2011, Eckhardt et al. 2006). Des travaux ont également montré que la méthylation asymétrique de cytosines dans les promoteurs joue un rôle dans la régulation de l'expression de quelques gènes, dans les cellules souches embryonnaires et les cellules somatiques où elle est présente (Pinney 2014). La méthylation du corps des gènes, quant à elle, est d'autant plus forte que le gènes est exprimé. Si elle ne bloque pas l'expression des gènes, elle joue probablement un rôle dans la répression de l'expression des éléments répétés intorniques et des promoteurs alternatifs (Jones 2012, Schübeler 2015). Enfin, l'affinité de certains facteurs de transcription pour l'ADN dépend de l'état de méthylation des cytosines présentes dans leurs sites de fixation (Hu et al. 2013) révélant une interaction potentiellement complexe entre méthylation de l'ADN,

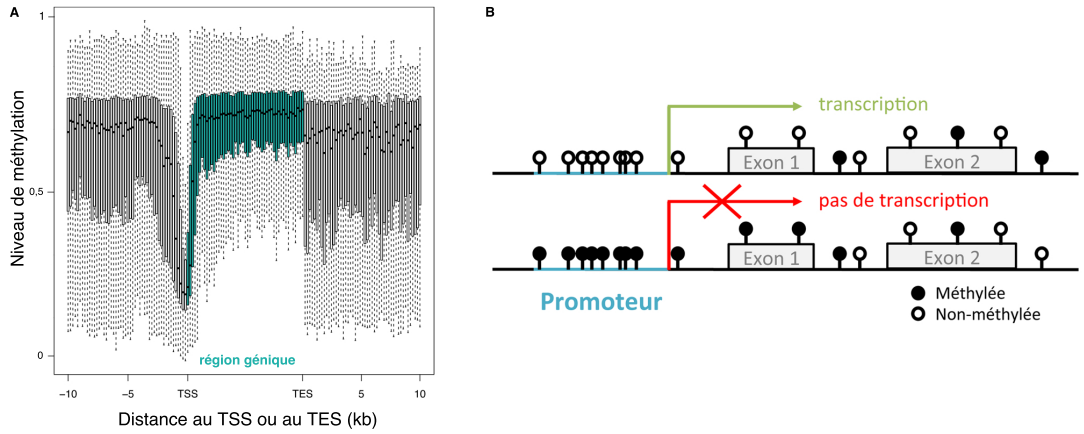


Fig. 10 Méthylation des promoteurs et expression des gènes. (**A**) Profil de méthylation autour d'un gène. (**B**) Corrélation entre méthylation du promoteur et expression des gènes.

Chapitre 5

Variation des profils de méthylation de l'ADN et influence de divers facteurs

L'association de la méthylation de l'ADN avec le niveau de compaction de la chromatine et d'activité des facteurs de transcription en font un excellent marqueur pour accéder à l'état épigénétique d'un tissu (Burger et al. 2013), et l'étude de ses variations pourrait permettre d'en apprendre plus les variations phénotypiques entre individus. L'état des sites de méthylation dans le génome humain peut être détecté en utilisant différentes méthodes, toutes basées sur la transformation par le bisulfite qui convertit les C non méthylés en U. Ils sont alors lus comme des T, alors que les 5-mC et les 5-hmC ne sont pas affectés, et sont lus comme des C (Laird 2010, Schübeler 2015). Ensuite, plusieurs solutions sont possibles : le séquençage reste la méthode la plus résolutive et la méthode de référence pour établir le profil de méthylation d'un échantillon mais ne permet pas de faire la différence entre les 5-hmC et les 5-mC. Il est possible de réaliser, avant le séquençage, une immuno-précipitation ciblant les 5-mC (MeDIP, *methylated DNA immunoprecipitation*), qui permet de les différencier des 5-hmC. Il existe aussi diverses puces, dont la plus complète est Illumina Infinium Human Methylation 450K, qui mesure le niveau de méthylation à plus de 480,000 sites répartis sur l'ensemble du génome, et enrichis en sites localisés dans les îlots CpG et les promoteurs.

Nous avons déjà vu que les profils de méthylation varient entre les types cellulaires, notamment à cause de la répression stable d'un certain nombre de promoteurs au cours de la différenciation cellulaire et des différentes étapes du développement (Lister et al. 2009, Smith and Meissner 2013). Ces différences

systématiques du niveau de méthylation de certains sites entre types cellulaires peuvent même utilisés pour prédire la proportion respective de ces différents types cellulaires dans un tissu complexe comme le sang (Houseman et al. 2012, Koestler et al. 2013). De nombreuses études ont également porté sur la comparaison des profils de méthylation entre cellules saines et cancéreuses. Elles montrent d'importantes modifications des niveaux de méthylation à de nombreux sites, incluant notamment une hypométhylation généralisée et une hyperméthylation des promoteurs de certains gènes suppresseurs de tumeurs. Ces variations de méthylation sont accompagnée de modifications d'autres marques épigénétiques (modifications résumées dans Hattori and Ushijima (2014)). Certains sites de méthylation peuvent donc être utilisés comme marqueurs pour la détection des cancers et l'établissement d'un pronostic (Barrow and Michels 2014). Enfin, ces travaux ont permis de développer de nouveaux traitements anti-cancéreux basés sur des substances hypométhylantes (Peedicayil 2012). La comparaison des méthylomes a également permis de découvrir des variations de

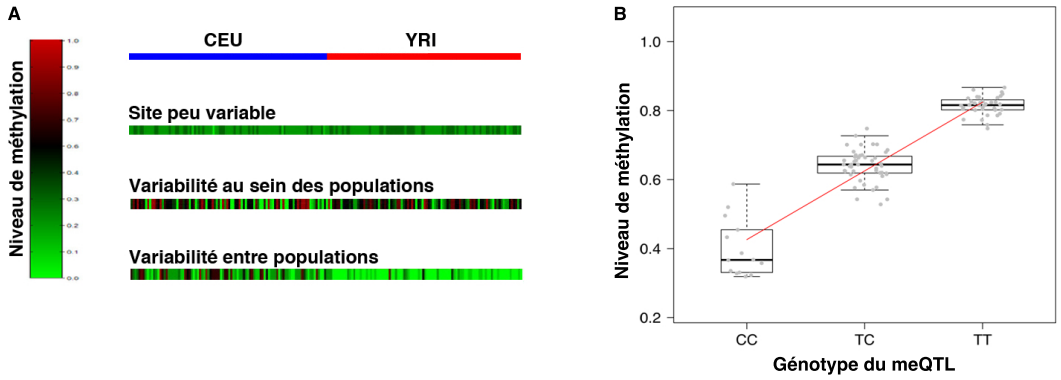


Fig. 11 Variations des profils de méthylation et facteurs génétiques. (**A**) Variations des profils de méthylation entre individus. Figure adaptée de Fraser et al. (2012). Mesures du niveau de méthylation de l'ADN à 3 sites CpG dans des lignées lymphoblastoïdes pour 90 Américains d'origine européenne (CEU) et 90 Yoruba du Nigéria (YRI). Les niveaux de méthylation sont indiqués par une échelle de couleur, vert signifiant une absence totale de méthylation, et rouge, une méthylation complète. (**B**) Exemple d'un meQTL. Le niveau de méthylation est relié au génotype d'un SNP voisin.

5.1 Les variations des profils de méthylation

5.1.1 Variabilité des profils de méthylation au cours de la vie

Les profils de méthylation peuvent varier chez un individu, pour un type cellulaire donné, au cours de sa vie. On observe en effet une hypométhylation globale du génome et une hyperméthylation des îlots CpG avec l'âge (Bjornsson et al. 2008, Boks et al. 2009, Christensen et al. 2009). De plus, des études d'association entre méthylome et âge ont permis de dresser des listes de sites pouvant servir de marqueurs de l'âge dans différents tissus (Bell et al. 2012, Hannum et al. 2013). Même si une partie de ces sites présentent des associations spécifiques d'un tissu, un certain nombre de sites, en particulier dans des gènes liés au développement (Rakyan et al. 2011), varient de la même façon avec l'âge dans de nombreux tissus. Ils constituent d'excellent biomarqueurs de l'âge et plusieurs algorithmes les utilisent afin d'estimer l'âge « épigénétique » d'une personne à partir de son méthylome (Hannum et al. 2013, Horvath 2013, Weidner et al. 2014)). On parle alors d'horloge épigénétique. Ces variations du profil de méthylation pourraient jouer un rôle dans le processus de vieillissement (Christensen et al. 2009, Murgatroyd et al. 2010). La vitesse de celle-ci peut être modifiée dans certains cas : par exemple l'obésité est associée à son accélération dans le foie (Horvath et al. 2014), le diabète de type 2 à son accélération dans les leucocytes (Toperoff et al. 2015) et les cellules cancéreuses montrent également des marques similaires à celles observées lors du vieillissement (Horvath 2013).

5.1.2 Variabilité des profils de méthylation entre individus et entre populations

Même si des variations de niveau de méthylation de l'ADN entre différents individus d'une même espèce ou entre représentants d'espèces différentes (par exemple homme et chimpanzé) pour un même tissu sont moins importantes que les variations entre tissus au sein d'un individu (Pai et al. 2011), de plus en plus de travaux se concentrent sur les associations entre les variations du méthylome et certains traits phénotypiques à l'échelle du génome (EWAS, *epigenetic whole-genome*

association studies, Murphy and Mill (2014), Rakyan et al. (2011)). Un petit nombre d'études ont également relevé des différences globales de niveau de méthylation entre populations à certains sites (Bell et al. 2011, Fraser et al. 2012, Heyn et al. 2013, Moen et al. 2013). De manière générale, un certain nombre de paramètres doivent être attentivement considérés lors de la réalisation d'études comparant les niveaux de méthylation entre individus ou populations. Différents types cellulaires présentent différents méthylomes qui sont étroitement liés à leurs fonctions. Cette observation a deux conséquences majeures. D'abord, le choix du type cellulaire doit se faire en fonction du phénotype étudié, même s'il est possible d'observer des modifications épigénétiques spécifiques dans les leucocytes chez les patients atteints de schizophrénie par exemple (Aberg et al. 2014). Ensuite, si le tissu étudié est composé de plusieurs types cellulaires, il est indispensable de corriger pour les variations de composition du tissu entre les différents individus

Les maladies auto-immunes tiennent une place particulière dans ce champ de recherche car des études chez des jumeaux monozygotes ont montré que l'héritabilité du risque de développer de telles maladies est très variable (12–67%) (Dang et al. 2013, Selmi et al. 2012). Elles ont également suggéré un rôle non négligeable des facteurs épigénétiques. Dans des cas de lupus érythémateux, d'arthrite rhumatoïde, de sclérose en plaque et de diabète de type 1 entre autre, on observe ainsi des modifications du niveau de méthylation lymphocytaire dans les promoteurs de certains gènes impliqués dans la reconnaissance du soi, l'immunité et l'inflammation (résumé dans Picascia et al. (2015)). D'autres études ont montré une association entre variations du méthylome et différents traits métaboliques, comme l'indice de masse corporelle, (Murphy and Mill 2014), la sensibilité à la douleur (Bell et al. 2014) et la schizophrénie (Aberg et al. 2014). De façon intéressante, des différences de méthylomes ont également été associées à des différences de réponses à certains traitements, notamment anti-cancéreux. En effet, l'un des enjeux majeurs en oncologie est de trouver des traitements fonctionnant chez le plus de patients possibles. Or les études cliniques ont montré une forte variabilité inter-individuelle des réponses aux médicaments (Tang et al. 2014). De plus en plus d'études se sont donc intéressées aux facteurs épigénétiques impliqués dans ces différences et ont montré des associations entre niveau de méthylation des promoteurs de gènes impliqués dans le transport et le métabolisme des molécules thérapeutiques et dans les voies de signalisation en

aval de leurs cibles (Tang et al. 2014). Toutes ces études vont donc en direction d'un lien fonctionnel entre diversité phénotypique et variations du méthylome. Cependant, la nature du lien (direct, indirect et impliquant l'intervention d'autres facteurs épigénétiques), et sa causalité (les modifications épigénétiques précèdent-elles ou sont-elles la conséquence des phénotypes observés ?) restent inconnues.

5.2 Variations des profils de méthylation : facteurs génétiques et environnementaux

5.2.1 Les facteurs génétiques de la variabilité des profils de méthylation

Nous avons vu qu'il existait des variations entre individus et populations en terme de profil de méthylation. En plus d'éventuelles variations stochastiques, deux facteurs peuvent expliquer les différences observées : les facteurs génétiques et les facteurs environnementaux, incluant les facteurs sociaux (Majnik and Lane 2014). Les études comparant des jumeaux monozygotes et dizygotes ont montré que les profils de méthylation de l'ADN des premiers étaient plus corrélés que ceux des seconds, suggérant l'existence de facteurs génétiques ayant une influence sur le niveau de méthylation (Ollikainen et al. 2010, Schneider et al. 2010). De même, diverses études de sites présentant un profil de méthylation allèle-spécifique chez des trios parents/enfant (Gertz et al. 2011), ou chez des individus non apparentés (Kerkel et al. 2008, Shoemaker et al. 2010) concluent à une grande influence des facteurs génétiques sur les profils de méthylation.

L'utilisation d'approches venant de la génomique et de la transcriptomique, qui consistent à corréler les niveaux de méthylation à certains sites avec les génotypes à des SNP voisins au sein d'une population ont permis l'identification de meQTL (*methylation quantitative trait loci*, figure 11B) dans divers tissus et populations (Bell et al. 2011, Fraser et al. 2012, Gutierrez-Arcelus et al. 2013, Heyn et al. 2013, Moen et al. 2013, Pai et al. 2015, Wagner et al. 2014). Si les meQTLs sont en général communs à différents tissus (Fraser et al. 2012, Smith et al. 2014), les comparaisons des listes de meQTLs entre différentes populations révèlent plus d'hétérogénéité. On distingue en effet deux classes de meQTL : ceux qui sont communs à plusieurs

populations et ceux, relativement nombreux, qui sont spécifiques d'une population (Fraser et al. 2012, Heyn et al. 2013, Moen et al. 2013), suggérant l'existence d'interactions entre plusieurs facteurs génétiques ou entre facteurs génétiques et environnementaux et niveaux de méthylation.

Plusieurs mécanismes ont été proposés pour expliquer l'association entre mutations génétiques et variations du niveau de méthylation. Premièrement, lorsque la mutation provoque la disruption d'un site CpG et donc la disparition de la méthylation des C, on observe une diminution globale du niveau de méthylation à très courte distance (40-50pb, Zhi et al. (2013)). Deuxièmement, certaines mutations affectant des éléments régulateurs (activateurs, promoteurs) pourraient modifier l'affinité de facteurs de transcription pour leur site de fixation, provoquant des modifications de la structure de la chromatine associées à des modifications de la méthylation de l'ADN (Kasowski et al. 2010). Une étude récente a d'ailleurs noté un enrichissement des sites de fixations en meQTL (Kaplow et al. 2015). Plus généralement, les variations du niveau de méthylation à des sites associés à des meQTLs semblent donc être la conséquence, directe ou indirecte, des mutations (Heyn 2014). Cependant, il semble que les meQTL ne soient que rarement également associés au niveau d'expression de gènes (Gibbs et al. 2010, Grundberg et al. 2012, Zhang et al. 2010), et que lorsque c'est le cas, les liens entre mutation, méthylation de l'ADN et niveau d'expression sont complexes. La méthylation peut en effet être selon les cas, un intermédiaire entre génétique et expression, une conséquence des modifications du niveau d'expression, ou bien en être complètement indépendante (Gutierrez-Arcelus et al. 2013).

5.2.2 Les facteurs environnementaux de la variabilité des profils de méthylation

De nombreux résultats plaident en faveur d'un impact des facteurs environnementaux sur le profil épigénétique en général, et la méthylation de l'ADN en particulier. Si la comparaison des profils de méthylation de jumeaux monozygotes à la naissance montre qu'il existe des différences de méthylation au sein des paires (Ollikainen et al. 2010, Schneider et al. 2010), il est difficile de faire la différence entre influence des facteurs environnementaux et variations stochastiques, les jumeaux partageant la plupart du temps des environnements très semblables. Cependant, le nombre de différences au sein de paires de jumeaux monozygotes semble augmenter avec le

temps, et ce indépendamment des modifications liées à l'âge (Fraga et al. 2005). De façon intéressante, le temps passé dans des environnements différents et l'ancienneté de la discordance des modes de vie sont corrélés au nombre de différences de méthylation observées, mettant en lumière le rôle de l'environnement dans ces variations du méthylome.

De nombreux facteurs environnementaux ont ainsi la capacité de modifier le profil de méthylation. Ainsi, l'exposition au soleil provoque des changements de niveau de méthylation des cellules de l'épiderme et du derme. Les modifications observées sont semblables à celles observées dans le cancer du colon et de la peau, suggérant un lien entre dommages épigénétiques dus à une exposition au soleil et conséquences phénotypiques (Vandiver et al. 2015). D'autres facteurs comme le fait de fumer entraîne des modifications de niveau méthylation de l'ADN (Tsaprouni et al. 2014). De façon intéressante, ces changements induits chez l'adulte sont partiellement réversibles après arrêt de la cigarette. Au contraire, l'exposition foetale à la fumée de cigarette semble, elle, provoquer des modifications à long terme du méthylome (Toledo-Rodriguez et al. 2010). Les modifications épigénétiques d'origine environnementale lors du développement embryonnaire et post-embryonnaire présentent donc un intérêt particulier, car elles sont susceptibles d'affecter un individu tout au long de sa vie.

On observe un certain nombre de modifications épigénétiques associées à l'exposition foetale à divers facteurs environnementaux (Teh et al. 2014) : modifications dans plusieurs tissus du niveau de méthylation dans le promoteur de certains gènes, dont notamment le facteur de croissance soumis à empreinte parentale *IGF2*, associés à un tabagisme actif ou passif de la mère durant la grossesse (résultats résumés dans Nielsen et al. (2014)), effets de la nutrition de la mère sur le méthylome de l'enfant (Heijmans et al. 2008, Tobi et al. 2009). Des expériences sur des souris ont montré que l'exposition à des perturbateurs endocriniens durant le développement embryonnaire et pré-pubertaire avait des conséquences sur le niveau de méthylation à plusieurs sites dans différents tissus (Skinner 2007). Des effets similaires ont été relevés chez l'Homme (LaRocca et al. 2014). Dans la plupart des cas, il semble cependant que ces effets ne soient pas héritables au cours des générations (Joubert et al. 2014).

Si les détails mécanistiques qui sous-tendent la modification des méthylomes par

les facteurs environnementaux ne sont pas connus, un lien commence à se dessiner entre exposition à certains facteurs environnementaux pendant le développement embryonnaire et l'enfance d'une part, et augmentation du risque de développer certaines pathologies d'autre part. Ainsi, l'exposition foetale à la famine entraîne des modifications de sites impliqués dans le vieillissement. Elle est également associée à l'augmentation du risque de développer des maladies comme des diabètes de types 2, des maladies cardiovasculaires et de syndromes métaboliques, qui sont des pathologies fortement liées au vieillissement (Tarry-Adkins and Ozanne 2014). Des études ont également montré l'impact du niveau socio-économique durant le développement embryonnaire et la petite enfance sur le niveau de méthylation de l'ADN à plusieurs sites modifications du méthylome (Lam et al. 2012). Plus généralement, le stress maternel, lors du développement embryonnaire, et parental, lors de la petite enfance, ont des effets sur le méthylome de l'enfant, effets qui persistent au moins jusqu'à l'adolescence (Essex et al. 2013). Des modifications du méthylome liées à un fort stress pendant les premiers jours de vie ont été également observés chez la souris, et sont accompagnés d'une plus forte expression de l'arginine vasopressine dans le noyau paraventriculaire de l'hypothalamus (Murgatroyd et al. 2009).

De plus en plus d'études montrent un lien chez l'Homme entre les modifications du méthylome par des facteurs environnementaux lors du développement embryonnaire et de la petite enfance et la variation de l'expression de certains gènes, par exemple chez les nouveaux-nés exposés à la cigarette pendant la grossesse (Nielsen et al. 2014, Suter et al. 2013), suggérant que l'épigénome peut servir d'interface pour intégrer les signaux environnementaux et modifier l'expression des gènes (Jaenisch and Bird 2003, Mazzio and Soliman 2012). De manière plus générale, l'accumulation des études sur les facteurs génétiques et environnementaux des variations épigénétiques souligne l'importance de considérer l'interaction entre ces différents acteurs pour comprendre les différences phénotypiques (Feinberg 2007, Sun 2014).

5.2.3 Héritabilité des profils de méthylation

Lorsque l'on parle d'héritabilité des profils de méthylation chez les mammifères en général, et dans l'espèce humaine en particulier, il faut distinguer l'héritabilité à travers les divisions cellulaires, clairement établie, et pour laquelle il existe des

mécanismes (cf. 4, Zaidi et al. (2010)), et l'héritabilité trans-générationelle, beaucoup plus discutée (Heard and Martienssen 2014). Si la découverte des meQTL propose un mécanisme pouvant assurer la transmission au cours des générations de variations du profil de méthylation causées par des mutations génétiques, la transmissions de modifications associées à des facteurs non génétiques apparaît moins probable. En effet, lors la gamétogenèse, puis de l'embryogenèse, les profils de méthylation sont ré-initialisés via deux déméthylations successives et quasiment complètes du génome (Seisenberger et al. 2012, Smallwood and Kelsey 2012), des étapes nécessaires à la programmation des cellules embryonnaires en cellules souches (Feng et al. 2010, Meissner 2010). Une étude de McRae et al. (2014) a estimé l'héritabilité de la méthylation à des sites non associés à des meQTL à 0,187.

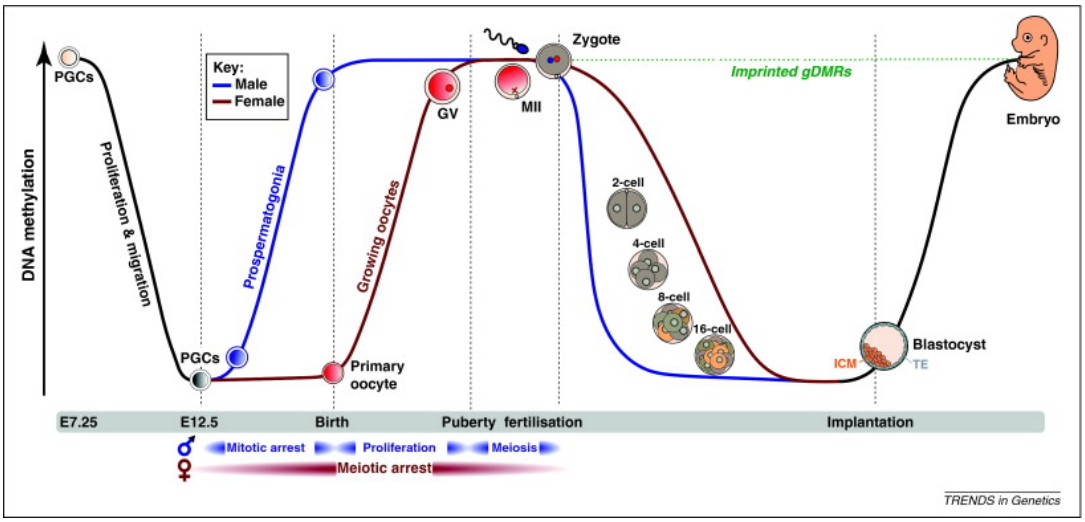


Fig. 12 Ré-initialisation des profils de méthylation de l'ADN pendant la gamétogenèse et l'embryogenèse. Figure tirée de Smallwood and Kelsey (2012). Les courbes représentent l'évolution du niveau global de méthylation de l'ADN au cours du développement. La courbe rouge correspond à l'ovogenèse, la bleue à la spermatogenèse, et la courbe noire aux cellules germinales primordiales (à gauche) et aux cellules embryonnaires (à droite). L'échelle est qualitative et correspond à un niveau global de méthylation très faible en bas, et à une méthylation de quasiment tous les sites CpG en haut.

Une autre difficulté pour détecter la transmission des marques épigénétiques est le fait que l'exposition d'une femme enceinte à des facteurs environnementaux va provoquer des modifications du profil de méthylation sur trois générations par exposition directe : sur la mère, l'enfant à naître, et la descendance de celui-ci, via des modifications sur les cellules germinales en cours de différenciation (Heard and Martienssen 2014). De même l'exposition d'un individu va provoquer une modification directe de l'épigénome sur deux générations via l'exposition des

ovocytes et des spermatozoïdes, sans « héritage » à proprement parler. Il faut donc observer la génération F2 ou F3 pour étudier des phénomènes d'héritage, et ils sont alors très rares (Heard and Martienssen 2014). Malgré ces fortes objections, des études chez la souris ont montré la possibilité d'hériter de traits phénotypiques via d'autres marques épigénétiques comme les ARN, notamment les petits ARNi, très présents dans les ovocytes et les spermatozoïdes, et qui sont transmis à la cellule-oeuf (Jablonka and Raz 2009). Il existe ainsi chez la souris un exemple de transmission d'un phénotype particulier de coloration de la queue et des pattes d'un mâle à sa descendance sans transmission de la mutation causale, qui se produit via des microARN, une classe particulière d'ARNi.

Pour résumer, il existe donc des variations des profils de méthylation entre individus et entre populations. Ces différences peuvent être dues à des facteurs génétiques et être transmises au cours des générations, ou à des facteurs environnementaux, et dans ce cas leur transmission à la génération suivante est très rare. Si de nombreuses variations inter-individuelles semblent être liées à des facteurs génétiques, il n'existe pas d'estimation de l'effet des facteurs environnementaux dans leur globalité sur les variations des profils épigénétiques.

OBJECTIFS DE LA THÈSE

Pour mieux comprendre l'effet de l'environnement sur la diversité phénotypique humaine, il apparait nécessaire de s'intéresser d'abord au rôle des facteurs environnementaux dans la diversité génétique et la diversité épigénétique des populations humaines. Comme nous l'avons vu dans l'introduction, l'environnement agit sur la diversité phénotypique à long terme des populations humaines via un mécanisme ciblant des mutations génétiques, la sélection naturelle. S'il semble qu'une majorité de gènes évoluent sous sélection négative chez l'Homme, différents travaux ont permis d'établir qu'un certain nombre de régions génomiques évoluent sous sélection positive. Or la mesure dans laquelle la sélection positive en général, et les balayages sélectifs en particulier, ont modelé la diversité génétique humaine est un sujet de controverse.

En outre, l'étude de la diversité génétique paraît de plus en plus insuffisante pour expliquer l'ensemble de la diversité phénotypique observée. L'épigénétique, qui participe à la régulation de l'expression des gènes sans changer l'information génétique, semble être un candidat intéressant pour expliquer une autre partie de la variabilité phénotypique humaine. Nous avons également vu précédemment que les facteurs environnementaux pouvaient modifier rapidement des profiles épigénétiques à certains sites chez les mammifères, et donc avoir un potentiel impact sur le phénotype à court terme. Il n'existe cependant pas d'évaluation de l'effet de l'environnement sur le profile épigénétique au niveau plus global du génome entier. Afin d'obtenir une image plus précise de l'ampleur de l'impact de l'environnement sur le génome et l'épigénome humain, mon travail de recherche s'est concentré sur deux principaux objectifs.

Premièrement, nous nous sommes concentrés sur l'étude des événements de sélection positive, et plus particulièrement des balayages sélectifs à l'échelle du génome en utilisant des données provenant de séquençage génome entiers. Pour cela, nous avons évalué la puissance de détection et la robustesse des différentes statistiques existantes dans le contexte spécifique des données de séquençage de nouvelle génération. Nous avons ensuite appliqué ces statistiques à des données de séquençage génome entier issues de bases de données publiques comme celle du *1000 Genome Project* ou de *Complete Genomics*, afin de déterminer s'il existe des événements de balayage sélectif récents, quel a été leur impact sur l'évolution récente du génome humain et leur rôle dans un certain nombre de traits phénotypiques et

maladies.

Deuxièmement, nous avons étudié l'impact de l'environnement sur les profiles de méthylation de l'ADN de plusieurs populations d'agriculteurs et de chasseurs-cueilleurs d'Afrique Centrale. Le fait que la méthylation de l'ADN soit le marqueur épigénétique le plus facilement accessible, et qu'il reflète les variations du paysage épigénétique global a motivé le choix du modèle. Ce projet s'est déroulé autour de deux axes : d'une part, mesurer l'effet respectif de l'habitat récent et des facteurs environnementaux et génétiques anciens sur les variations des profils de méthylation de l'ADN de populations afin de déterminer les sites et fonctions biologiques affectés par chacun et d'autre part, étudier la proportion de variation des profils de méthylation imputable à des facteurs génétiques afin d'évaluer l'impact relatif des facteurs environnementaux et génétiques sur la diversité des profiles de méthylation de l'ADN.

L'objectif est donc d'obtenir une image globale de l'effet de l'environnement sur la variabilité du génome et l'épigénome humain. Cela pourrait ainsi permettre de mieux appréhender le rôle joué par l'environnement sur la diversité phénotypique humaine actuelle, et d'ouvrir la voie à une réflexion sur les modes d'adaptation de l'Homme à son environnement.

RÉSULTATS

Chapitre 6

Existence et fréquence des balayages sélectifs dans le génome humain

6.1 Contexte

Au cours de ces dernières années, de nombreuses études de sélection positive au niveau génomique ont mené à l'identification de plusieurs centaines de régions portant des signatures d'évolution adaptative dans différentes populations humaines. Ces travaux, en montrant que certaines fonctions biologiques ont été particulièrement ciblées par la sélection positive, ont permis de mieux comprendre la manière dont l'environnement a agi sur la diversité humaine et ont permis de caractériser une partie des pressions sélectives ayant guidé l'évolution du génome humain. Cependant, comme nous l'avons vu dans l'introduction, l'ampleur de l'effet de la sélection positive sur la diversité du génome humain est sujet à controverse. En effet, certaines études affirment que la sélection positive a joué un rôle majeur dans l'évolution du génome humain, alors que d'autres suggèrent que les signaux de sélection détectés pourraient être le résultat d'un certains nombres de facteurs confondant. En particulier, de nouvelles études affirment qu'il se serait produit très peu d'événements récents de balayage sélectif lors l'évolution du génome humain.

Les récentes données de séquençage « génome entier » d'un certain nombre d'individus issus de plusieurs populations ont donné accès à un nombre important de SNP et ont permis de supprimer un certain nombre de biais, laissant espérer une augmentation de la puissance des méthodes de détection des balayages sélectifs. Cependant, leur analyse n'a pas permis jusqu'à présent de résoudre les contradictions observées entre les différents travaux. Il devient donc nécessaire, avant toute étude

d'évaluation de la prévalence de la sélection dans le génome humain, de mesurer la puissance des statistiques existantes à détecter la sélection positive dans le contexte précis des données de séquençage et leur robustesse à un certain nombre de facteurs confondants. En effet, de nombreux jeux de données, dont une partie de *The 1 000 Genomes Project*, présentent une profondeur de séquençage assez faible, une caractéristique susceptible d'affecter la puissance des statistiques (Crawford and Lazzaro 2012). De plus, ceux avec une grande profondeur, tels que *Complete Genomics* ont généralement une taille d'échantillon assez faible, autre paramètre pouvant diminuer la puissance de détection.

Notre étude s'est donc d'abord attachée à tester l'effet d'une faible profondeur de couverture ou d'un petit nombre d'échantillons sur la puissance de détection de statistiques permettant de détecter des événements de balayage sélectif dans le contexte d'une histoire démographique humaine réaliste. Après avoir vérifié que les statistiques utilisées n'étaient sensibles ni aux autres modes de sélection positive alternatifs tels que l'adaptation polygénique ou la sélection sur variant pré-existant ni à sélection d'arrière plan, nous les avons appliqués aux données de *The 1 000 Genomes Project* et de *Complete Genomics*, afin de déterminer si des événements de balayages sélectifs s'étaient produits récemment au cours de l'évolution des population humaines.

6.2 Article 1 : Exploring the occurence of classic selective sweeps in humans using whole-genome sequencing data sets

Exploring the Occurrence of Classic Selective Sweeps in Humans Using Whole-Genome Sequencing Data Sets

Maud Fagny,[1,2,3] Etienne Patin,[1,2] David Enard,[4] Luis B. Barreiro,[5] Lluis Quintana-Murci,*[,1,2] and Guillaume Laval*[,1,2]

[1]Institut Pasteur, Human Evolutionary Genetics, Department of Genomes and Genetics, Paris, France

[2]Centre National de la Recherche Scientifique, URA3012, Paris, France

[3]Université Pierre et Marie Curie, Cellule Pasteur UPMC, Paris, France

[4]Department of Biology, Stanford University

[5]Department of Pediatrics, Sainte-Justine Hospital Research Center, University of Montreal, Montreal, Quebec, Canada

***Corresponding author:** E-mail: glaval@pasteur.fr; quintana@pasteur.fr.

Associate editor: Ryan Hernandez

Abstract

Genome-wide scans for selection have identified multiple regions of the human genome as being targeted by positive selection. However, only a small proportion has been replicated across studies, and the prevalence of positive selection as a mechanism of adaptive change in humans remains controversial. Here we explore the power of two haplotype-based statistics—the integrated haplotype score (iHS) and the Derived Intraallelic Nucleotide Diversity (DIND) test—in the context of next-generation sequencing data, and evaluate their robustness to demography and other selection modes. We show that these statistics are both powerful for the detection of recent positive selection, regardless of population history, and robust to variation in coverage, with DIND being insensitive to very low coverage. We apply these statistics to whole-genome sequence data sets from the 1000 Genomes Project and Complete Genomics. We found that putative targets of selection were highly significantly enriched in genic and nonsynonymous single nucleotide polymorphisms, and that DIND was more powerful than iHS in the context of small sample sizes, low-quality genotype calling, or poor coverage. As we excluded genomic confounders and alternative selection models, such as background selection, the observed enrichment attests to the action of recent, strong positive selection. Further support to the adaptive significance of these genomic regions came from their enrichment in functional variants detected by genome-wide association studies, informing the relationship between past selection and current benign and disease-related phenotypic variation. Our results indicate that hard sweeps targeting low-frequency standing variation have played a moderate, albeit significant, role in recent human evolution.

Key words: **positive selection, whole-genome sequence data, human populations, neutrality statistics.**

Article

Introduction

The detection of genomic regions that have been targeted by recent positive selection has proved a powerful tool for delineating genes contributing to adaptation to environmental variables and for informing functions accounting for phenotypic diversity. Over the last decade, many genome-wide scans for selection have been reported in humans, fueled by the advent of whole-genome single nucleotide polymorphism (SNP) data sets. These studies have made use of various statistical methods based on the predictable effects of positive selection on patterns of genetic variation. These effects include a decrease in haplotype diversity (Voight et al. 2006; Frazer et al. 2007; Sabeti et al. 2007; Tang et al. 2007; Pickrell et al. 2009), high fraction of rare alleles (Carlson et al. 2005; Kelley et al. 2006), or major shifts of allele frequency between populations (Akey et al. 2002; Hinds et al. 2005; Weir et al. 2005; Frazer et al. 2007; Barreiro et al. 2008; Chen et al. 2010; Oleksyk et al. 2010; Jin et al. 2012). These approaches have led to the identification of several hundred genomic regions displaying selection signals, suggesting the presence in these regions of new beneficial mutations that have spread rapidly through the population.

The more recent advent of whole-genome sequence (WGS) data sets has provided unbiased information relating to the spectrum of allelic variation, overcoming the SNP ascertainment biases that characterize SNP genotyping data sets, with a power of ~99% to detect variants with a population frequency above 1%, for most of the genome (Abecasis et al. 2012). For example, the 1000 Genomes (1000G) project, both its Pilot and Phase 1 releases (1000 Genomes Project Consortium 2010; Abecasis et al. 2012), and the Complete Genomics (CG) data set (Drmanac et al. 2010) have provided with 12–38 million SNPs from various populations worldwide. This dramatic increase in the amount of sequence information available, corresponding to up to ten times that provided by the HapMap Consortium (Frazer et al. 2007; Altshuler et al. 2010), should provide increased power for evaluating the impact and prevalence of selection on the human genome. In this context, a recent study of the 1000G Pilot data set has defined a list of genes for which there was compelling evidence of positive selection (Grossman et al. 2013).

Despite the considerable contribution of genome-wide scans to our understanding of the effects of natural selection on patterns of genome diversity, replication in different studies and functional support for adaptive significance have been demonstrated for only a handful of genes (Akey 2009). Furthermore, and more generally, the importance of positive selection in shaping human diversity remains an open question. Some studies have reported enrichment of certain functional SNP classes among selection signals, suggesting a nonnegligible prevalence of positive selection as a driving force of human adaptation (Voight et al. 2006; Frazer et al. 2007; Barreiro et al. 2008; 1000 Genomes Project Consortium 2010; Jin et al. 2012). However, others have suggested that these enrichment signals might actually result, at least in part, from the action of background selection (Coop et al. 2009; Pritchard et al. 2010; Hernandez et al. 2011). In addition, some studies indicate that selection following the "hard sweep" model, in which new advantageous mutations arise and spread rapidly to fixation, has occurred only rarely in recent human evolution (Hernandez et al. 2011; Granka et al. 2012). Indeed, it has been proposed that many adaptive events have occurred through other, largely undetected forms of positive selection, such as polygenic adaptation or selection on standing variation (Pritchard and Di Rienzo 2010; Pritchard et al. 2010).

The lack of agreement between these selection studies highlight the need to assess the power of statistical methods for detecting the effects of positive selection in the context of human demography and specifically of WGS (e.g., coverage, SNP calling, number of individuals). For example, simulations of populations of *Drosophila* and *Anopheles* mosquitoes (i.e., large populations with constant sizes of 10^6 individuals) have already shown that low coverage can potentially impact the power to detect selective sweeps (Crawford and Lazzaro 2012). It also remains unclear whether the evidence of positive selection—that is, enrichment of genic regions, as opposed to nongenic regions, among selection signals (Voight et al. 2006; Barreiro et al. 2008; Jin et al. 2012)—can be extended to WGS data sets and is robust to alternative selection scenarios, such as background selection. In light of the increasing amount of WGS data sets, there is a methodological need to address these issues as an indispensable prerequisite to explore the occurrence of selection in the genomes of humans and other species.

In this study, we aimed to explore the prevalence of recent, strong positive selection (i.e., the hard sweep model) in human adaptation, using WGS data sets. To do so, we first performed a simulation study based on realistic models of human demography and determined the power of relevant neutrality statistics for detecting recent population-specific positive selection, considering the features of current WGS data sets, such as differences in coverage and sample size. We next evaluated the sensitivity of these statistics to other selective regimes, such as polygenic adaptation, positive selection on standing variation and background selection. We then analyzed the 1000G and CG data sets and found enrichment of some functional SNP classes among selection signals, controlling explicitly for potential confounding factors. Lastly, we searched for functional support of the adaptive significance of genomic regions enriched in selection signals and found that these regions are indeed enriched for SNPs associated with phenotypic variation, both benign and disease related.

Results

Power to Detect Recent Hard Sweeps from Next-Generation Sequencing Data

We first evaluated the power to detect recent hard sweeps over a large range of allele frequencies, from next-generation sequencing data. We simulated autosomal regions under neutral and hard sweep assumptions, using for both the same calibrated model designed to match realistic scenarios of human demography (fig. 1*A*; supplementary text, table S1, and fig. S1, Supplementary Material online) (Voight et al. 2005; Laval et al. 2010; Gravel et al. 2011). Indeed, publicly available WGS data sets, such as the 1000G and CG data sets, include continental populations with different demographic histories, a feature known to affect the power of neutrality statistics (Pickrell et al. 2009; Li 2011). We focused on two haplotype-based statistics that are known to exhibit high power to detect positive selection over a large range of allele frequencies (Voight et al. 2006; Barreiro et al. 2009) and expected to be insensitive to background selection, that is, there is no prior reason that background selection differentially affects the haplotypes sharing the ancestral or derived allele. This contrast with statistics based on population differentiation, such as F_{ST}, where distinguishing the effects of positive and background selection is more challenging (Hernandez et al. 2011). We thus used the integrated haplotype score (iHS), which measures the difference in haplotype homozygosity associated with the ancestral and derived alleles (Voight et al. 2006), and the Derived Intraallelic Nucleotide Diversity (DIND) test, which measures the differences in nucleotide diversity associated with the ancestral and derived alleles (Barreiro et al. 2009). This choice was also based on the fact that iHS has been successfully used to detect strong signals of positive selection in genotyping data—that is, significant enrichment of functional sites among selection signals (Voight et al. 2006), and DIND was designed to make full use of resequencing data (Barreiro et al. 2009). Furthermore, in line with our aims, both iHS and DIND exhibit substantial power over a large range of allele frequencies of the selected mutation (Voight et al. 2006; Barreiro et al. 2008), in contrast with other statistics such as XP-EHH or composite likelihood ratio (CLR), which are known to detect almost-completed or recently completed sweeps (i.e., frequency of the selected allele > 0.8) (Nielsen et al. 2005; Sabeti et al. 2007; Williamson et al. 2007; Casto et al. 2010).

To validate our simulation process, we estimated the power of iHS and DIND, assuming genotypes and gametic phases to be known (i.e., "full sequence data," fig. 1*B*, see Materials and Methods). We set $2Ns$ (N being the effective size and s being the selection coefficient) to 100 and simulated 100-kb DNA regions. Consistent with simpler scenarios of populations of constant size that used similar parameters

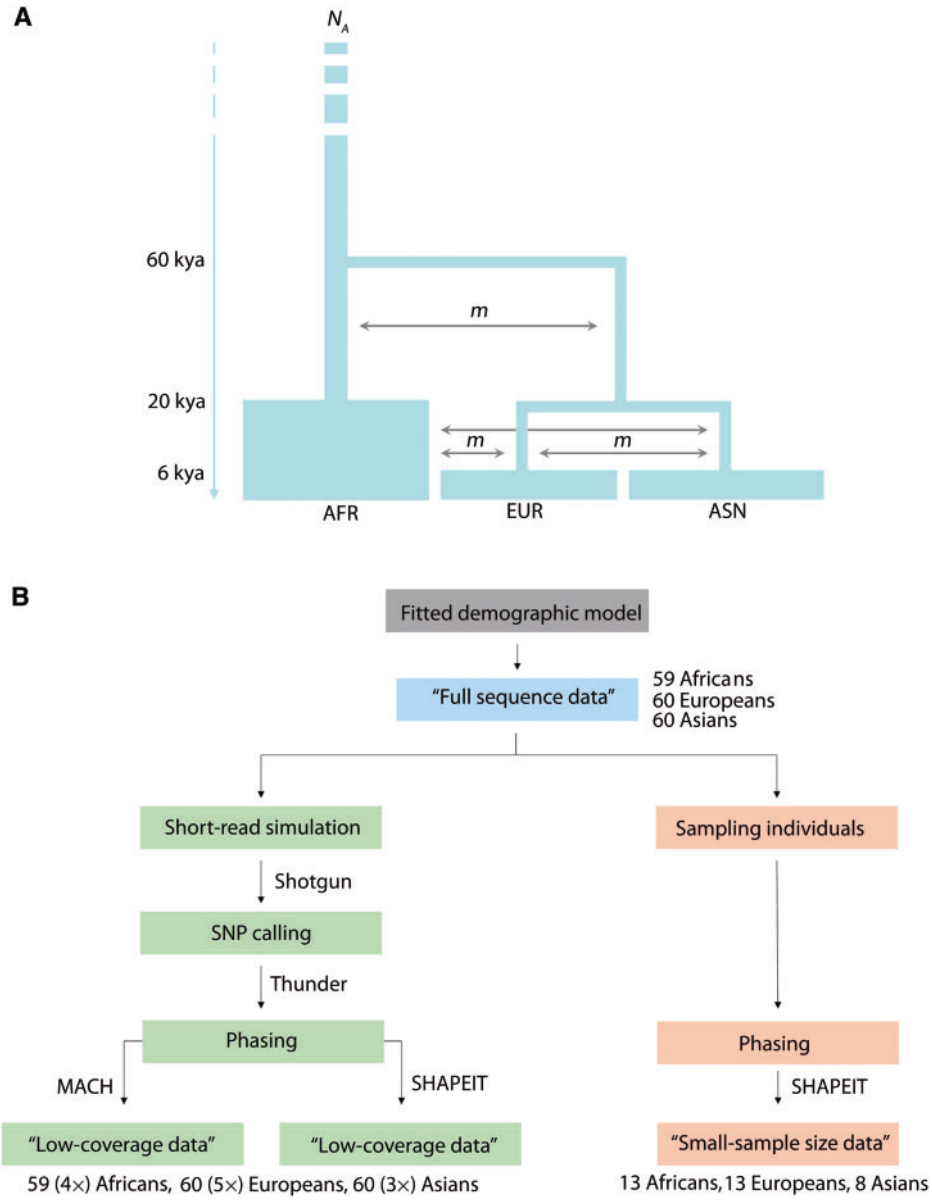


FIG. 1. Demographic model and flow chart used for the simulations. (*A*) Demographic model. The model used for the simulations considers that the ancestral Eurasian population split from the initial ancestral population (N_A) 60,000 years ago (60 kya) and went through a bottleneck reducing by half its effective population size. European and Asian populations diverged 20,000 years ago (20 kya) and went through recent expansions, corresponding to the Neolithic transition, increasing their effective size by 100. The expansion of the African population increased its effective size by 50. The migration parameter *m* is set to 1.3×10^{-5}. We used this calibrated demographic model to perform all subsequent simulations, that is, neutral simulations as well as those under various models of selection (recent selective sweep, background selection, and interaction between recent selective sweep and background selection, see Materials and Methods). (*B*) Flow chart for the simulations. To mimic 1000G Pilot data (green pipeline), we simulated low coverage from the "full sequence data" and inferred gametic phases. To mimic CG data (orange pipeline), we randomly sampled individuals from the "full sequence data" simulations and inferred gametic phases. Given the high coverage of the CG data set (read depth per site of 50× in average with 99% confidence interval ranging from 26× to 107× in Africa, 29× to 110× in Europe, and 17× to 75× in Asia), we did not simulate coverage, as this should not impact the power of DIND and iHS.

(Voight et al. 2006; Barreiro et al. 2009), iHS and DIND had a power of almost zero for selected allele frequencies (SAF) below 0.2, increasing rapidly to 80–100% for SAFs above 0.4 (fig. 2). In addition, the power computed as a function of SAF was similar to that found in a previous study specifying selection intensity on the basis of the age and the final frequency of the selected allele (Grossman et al. 2013). As expected, iHS and DIND clearly outperformed various neutrality statistics based on the allele frequency spectrum (AFS), even when we assumed realistic demographic models (supplementary table S2, Supplementary Material online). At the population level, the power of iHS and DIND was higher in African (78.90% and 76.06%, respectively) than in Eurasian populations (40.98% and 38.20%, respectively), highlighting the impact of demography on the power of these statistics. Furthermore, the power of both statistics was found to be similar after

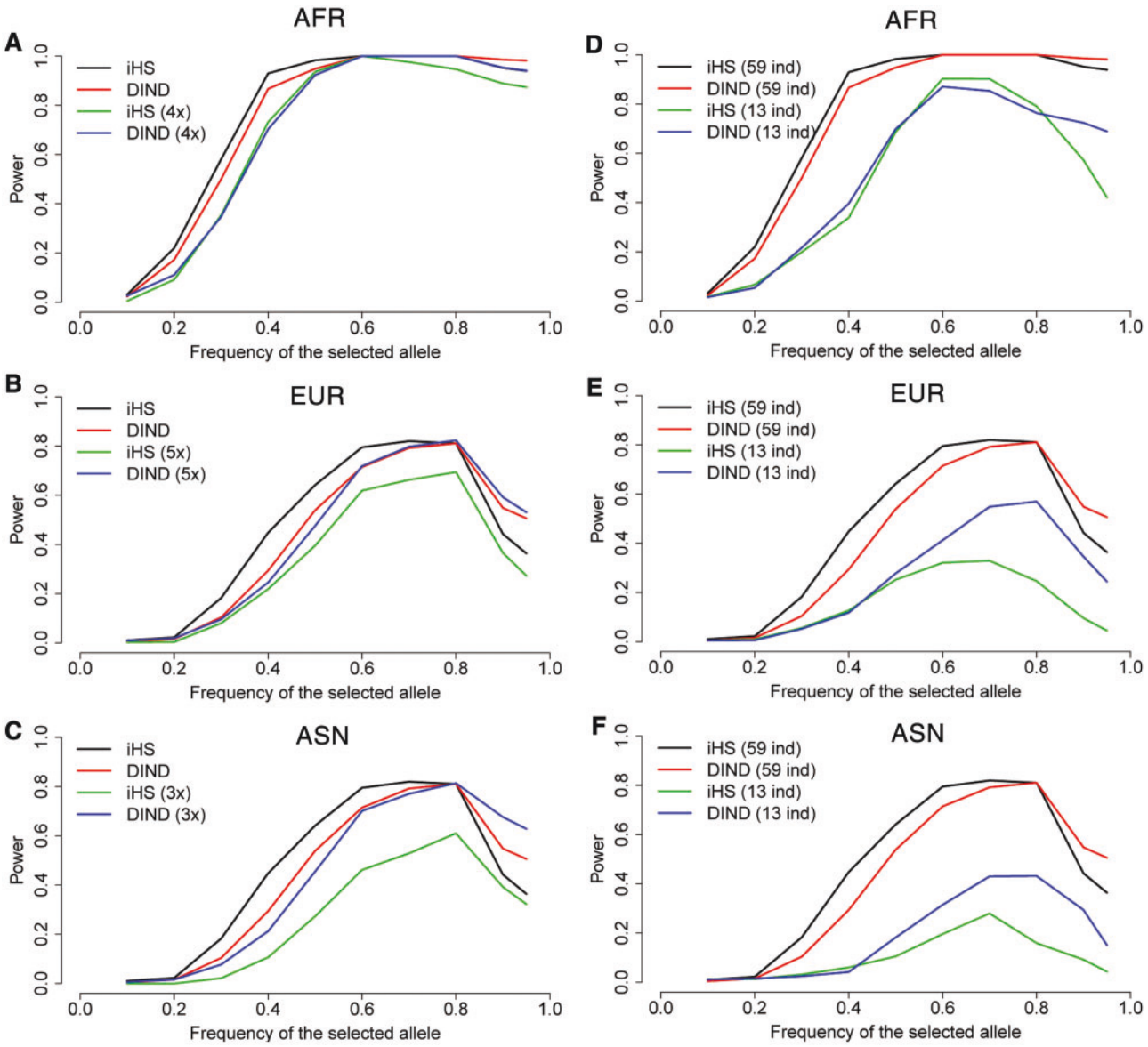


FIG. 2. Power of iHS and DIND to detect recent hard sweeps as a function of SAF. Critical values for both statistics, at FPR = 0.01, were obtained from 10^4 neutral simulations ($2Ns = 0$). For each simulation performed under recent positive selection ($2Ns = 100$), we used the proportion of extreme iHS and DIND values (see Materials and Methods). (*A*–*C*) Simulated "full sequence data" and "low-coverage data" (5× for Africans, 4× for Europeans, and 3× for Asians). (*D*–*F*) Simulated "full sequence data" and "small-sample size data" (13 individuals for Africans and Europeans, 8 individuals for Asians). In each case, we performed a total of about 2,000 simulations. (*A*, *D*) African population. (*B*, *E*) European population. (*C*, *F*) Asian population.

simulation of variation in recombination rate (i.e., presence of hotspots) and mutation rate (i.e., SNP density) (supplementary figs. S2 and S3, Supplementary Material online). The only exception to this trend was when SNP density was very low, where the power of iHS dropped dramatically as previously observed (Crisci et al. 2013), while that of DIND decreased only moderately. Overall, our simulations indicated that these haplotype-based statistics constituted powerful tests for detecting the effect of recent hard sweeps on a large range of allele frequencies in the context of full sequence data sets, regardless of the demographic history of the population considered.

We then investigated the effects of variation in coverage and sample size, characterizing WGS data sets, on the power of iHS and DIND. The 1000G Pilot data set is characterized by sample sizes of ~60 individuals per population sequenced at low coverage (3–5×), whereas the CG data set is characterized by small sample sizes (8–13 individuals per population) sequenced at high coverage (50×). We thus simulated data sets mimicking the 1000G Pilot ("low-coverage data") and CG ("small-sample size data") data sets and considered the uncertainty associated with haplotype phasing using MaCH and SHAPEIT (Li et al. 2010; Delaneau et al. 2012) (see Materials and Methods, fig. 1*B*). The power of both statistics varied with the frequency of the selected allele, as previously shown (fig. 2). Comparison of the full sequence and low-coverage simulated data sets demonstrated that low coverage had no impact on the power of DIND but slightly affected the

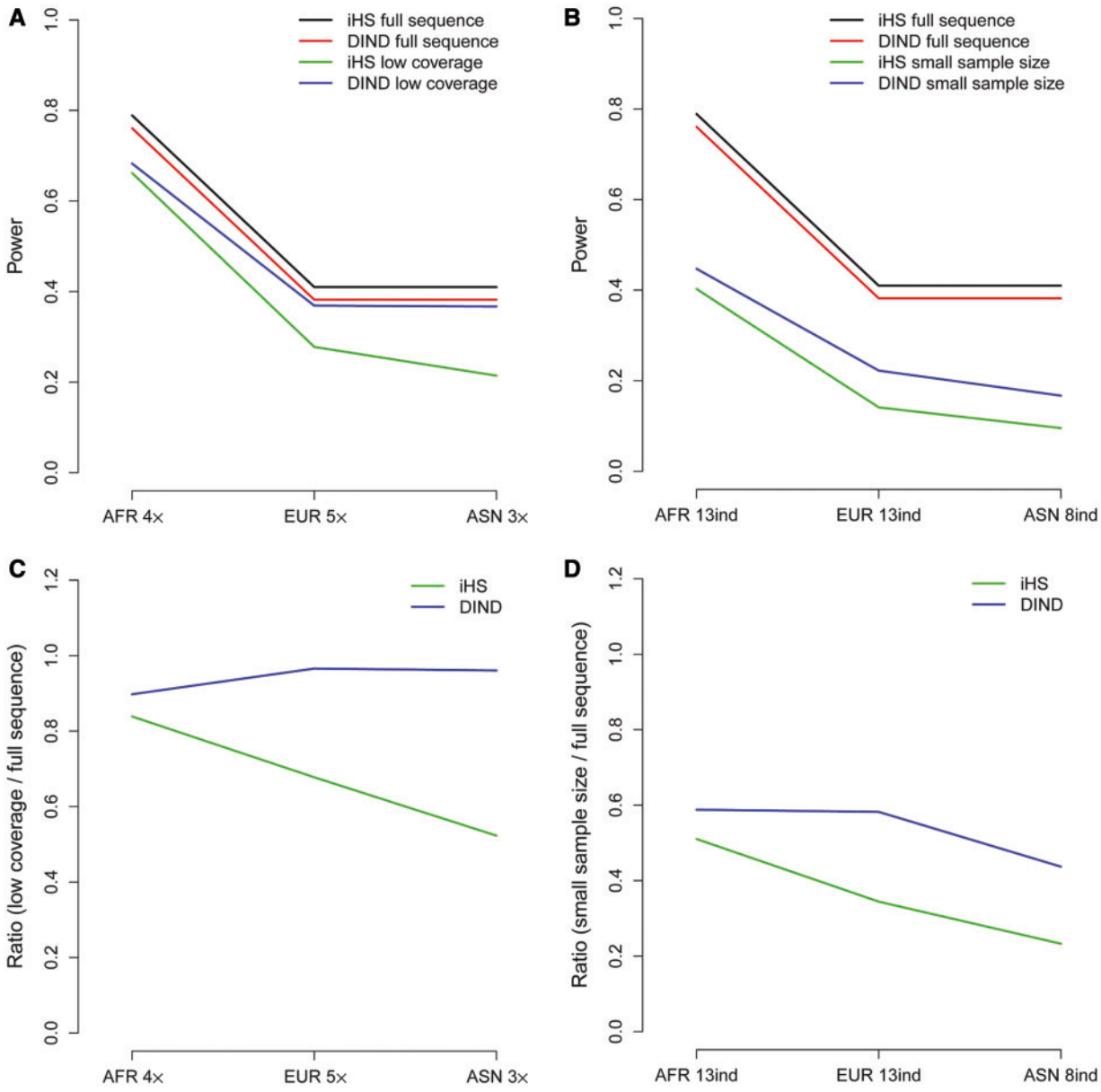


Fig. 3. Effect of low coverage and low sample size on the power to detect recent hard sweeps. (*A*) and (*B*) Power of iHS and DIND summed over a wide range of SAFs (simulations with SAF≥0.2 are considered together). (*C*) and (*D*) Power of iHS and DIND obtained within the context of next-generation sequencing data ("low-coverage data" or "small-sample size data") divided by the same power obtained with "full sequence data." For example, a ratio of 0.6 indicates that the power obtained with "low-coverage data" is 60% to that obtained with full sequence data. (*A*) and (*C*) "Low-coverage data" versus "full sequence data." The coverage is indicated for each population. (*B*) and (*D*) "Small-sample size data" versus "full sequence data" (60 individuals). The number of individuals is indicated for each population.

power of iHS (fig. 3*A* and *C*). By contrast, small sample sizes similar to those of the CG data set had a strong impact on the power of both statistics, this effect being most pronounced for iHS (fig. 3*B* and *D*). Note that the phasing process did not alter the power of iHS and DIND, as it was found to be similar with either inferred or known gametic phases (fig. 3*A*). In addition, the power was found to be similar when individual gametic phases were inferred either with MaCH or SHAPEIT (data not shown).

Overall, we found that sample size had a stronger effect on the power of these tests than coverage, which had little impact on power, with the DIND test being insensitive to even very low depth of coverage (~3×).

Robustness of iHS and DIND to Alternative Selective Regimes

Selective processes such as background selection, that is, the reduction in variability at neutral or nearly neutral sites due to selection against linked deleterious alleles (Charlesworth et al. 1997; Charlesworth 2012), can mimic the patterns left by positive selection, generating spurious "positive selection"

signals in some cases (Hernandez et al. 2011). We determined the extent to which iHS and DIND were sensitive to background selection. Given that 30–40% of human nonsynonymous mutations have been suggested to be highly deleterious or lethal ($|s| > 1\%$ i.e., $2Ns$ lower than -200 in humans, see Boyko et al. [2008]), we simulated genomic regions with 20% of sites under negative selection, mimicking the selective features that can be observed in coding regions. We used various values of the population genetic selection parameter $2Ns$, ranging from -1 to -500 (supplementary table S3, Supplementary Material online). The proportion of simulated sequences under background selection detected with iHS and DIND at a false positive rate (FPR) of 1% ranged from 0% to 2.5% (average ~1%), indicating that neither of these tests could detect this selective regime. Because the patterns of genetic variation can be the target of multiple modes of selection (Hernandez et al. 2011), we next explored whether background selection can alter the signal of a hard sweep. We tested whether negatively selected mutations segregating near positively selected variants affect the power of iHS and DIND, by simulating 100-kb regions in which a new advantageous mutation ($2Ns = 100$) was inserted in a genetic background where 20% of sites were negatively selected (see Materials and Methods). We found that background selection does not alter the power to detect selection following a hard sweep model (supplementary fig. S4, Supplementary Material online).

Alternative models of positive selection, such as polygenic adaptation or selection on standing variation, can also play an important role in adaptation, but their effects are more difficult to detect (Pritchard and Di Rienzo 2010; Pritchard et al. 2010). We evaluated the power of iHS and DIND to detect polygenic adaptation, which was modeled here as weak positive selection acting on many independent loci. This model of polygenic adaption has been proposed as an alternative model to rapid genetic adaptation, in light of the highly polygenic architecture of many traits in humans (Turchin et al. 2012). We thus simulated positive selection models with a low $2Ns$ ($2Ns = 5$, supplementary table S4, Supplementary Material online), keeping unchanged all the other parameters used to simulate hard sweeps (see Material and Methods). Neither DIND nor iHS detected a selection signal at low values of $2Ns$, as low $2Ns$ values lead to small shifts in the frequency of the selected alleles, as predicted under a model of polygenic adaptation acting through weak selection (Pritchard and Di Rienzo 2010; Pritchard et al. 2010). Consequently, our results support the notion that conventional methods have little power to detect signatures of polygenic adaptation (Chevin and Hospital 2008; Pritchard and Di Rienzo 2010; Pritchard et al. 2010).

Finally, we performed simulations of positive selection on standing variation, that is, a neutral or mildly deleterious allele that is already segregating in the population at a frequency greater than $1/2N$ suddenly becomes positively selected and increases in frequency (Przeworski et al. 2005; Pritchard and Di Rienzo 2010; Pritchard et al. 2010). We evaluated the power of iHS and DIND for an initial frequency of the selected allele from 0.01 to 0.5 and used values of $2Ns$ ranging from 100 to 1,000 (supplementary table S5, Supplementary Material online). To do so, we used mpop software (Pickrell et al. 2009), which allows simulations only in a constant-size population model (see Materials and Methods). The power of iHS and DIND was found to decrease with increasing initial frequency of the selected allele, as high initial frequencies reduce the signature of the sweep around the selected site (Przeworski et al. 2005). For example, the power of both statistics was lower, by a factor of 4, for initial frequencies of the selected allele ≥ 0.2 and a $2Ns = 100$. The application of such a decrease in power to the results of iHS and DIND obtained considering appropriate demographic histories and mimicking WGS data (fig. 3*A* and *B*) would yield a power of less than 10% for non-African samples. Moreover, no signals of positive selection on standing variation were detected (data not shown) when simulations were performed with low values of $2Ns$ ($2Ns < 10$), because the frequency shifts of the selected alleles were, as for polygenic adaptation by weak selection, too small to be detected.

Our simulation results demonstrate that iHS and DIND are insensitive to background selection and underpowered for the detection of polygenic adaptation or recent positive selection on standing variation when the selected allele has an initial frequency of 0.2 or above. Thus, the signals of positive selection detected by DIND and iHS in WGS data sets should reflect the effects of recent, strong positive selection targeting either a newly arisen allele (i.e., hard sweep *stricto sensu*) or standing mutations with a preselection frequency lower than 0.2 (nearly hard sweep).

Assessment of the Genome-Wide Extent of Selection Using Functional SNP Classes

To assess the extent of positive selection at the genome-wide level and to evaluate whether iHS and DIND are able to detect enrichment in selection signals in particular SNP functional classes from WGS data sets, we analyzed the 1000G and CG data sets (supplementary fig. S5 and table S6, Supplementary Material online). In classical outlier approaches, which identify SNPs presenting extreme values for a given statistic as displaying evidence of selection, the proportion of false positives remains unknown and can be high (Kelley et al. 2006; Teshima et al. 2006). Here, we overcome this caveat by applying the following rationale: if positive selection has preferentially targeted functionally important loci, then we would expect an enrichment of certain functional SNP classes among extreme values for a particular statistic (Voight et al. 2006; Barreiro et al. 2008; Jin et al. 2012). For example, it has been shown that positive selection can create strong clustering of extreme iHS values yielding strong enrichments of such extreme values within genes (Voight et al. 2006). We therefore investigated whether iHS and DIND outliers (the top 1% of values for each statistic) were more strongly enriched in putatively functional SNPs (i.e., genic or nonsynonymous SNPs) than in nongenic SNPs (supplementary table S7, Supplementary Material online). This approach, which allows quantifying the proportions of false-positive signals, should make it possible to

deduce the proportion of outliers genuinely targeted by positive selection.

We thus calculated iHS and DIND for the phased data of each population of the 1000G Pilot and the CG data sets, using windows of 100 kb centered on each SNP (i.e., the core SNP) and retaining only those for which the derived state of the core SNP was unambiguously determined. We minimized the FPR by excluding windows in which the core SNP had a derived allele frequency (DAF) below 0.2, given that these tests had a power close to zero in such conditions (fig. 2 and supplementary table S2, Supplementary Material online). We assessed enrichment of SNP classes among outliers by logistic regression, generating an odds ratio (OR) for the effect of recent positive selection. If selection has occurred in genic regions, an OR > 1 would be expected, reflecting the enrichment of genic SNPs among outliers (e.g., OR = 1.25 when there are 20% true and 80% false-positive SNPs among genic outliers). Otherwise (i.e., 100% of false positives among genic outliers), we would expect an OR ≤ 1, indicating that the proportion of genic SNPs among outliers is no greater than the proportion of genic SNPs among all SNPs (~38% for the 1000G and CG data sets, supplementary table S7, Supplementary Material online). We also controlled for various potential confounding factors, such as genomic variation in coverage, recombination rate, and the number of SNPs per window, and calculated corrected ORs (OR_C, see Materials and Methods).

With DIND, significant enrichment in genic SNPs was observed for both the 1000G Pilot and CG data sets (table 1). These enrichments were found to be robust to the confounding factors tested ($OR_C > 1$) and were highly significant when compared with several genomic resamplings (table 1, see Materials and Methods). Likewise, DIND outliers displayed a greater enrichment in nonsynonymous SNPs, with respect to nongenic SNPs, although the statistical significance of this enrichment was lower due to the small number of nonsynonymous SNPs tested (table 1). In contrast with the results obtained for Africans and Europeans, almost no significant enrichment was observed for Asians from the 1000G Pilot and CG data sets ($OR_C = 0.97$ and $OR_C = 0.98$, table 1). In the CG data set, sample size was the smallest for the Asian population, confirming the critical nature of this experimental specification (fig. 3*B* and *D*). In the 1000G Pilot data set, the low coverage of Asians (~3×) would not be expected to mask the enrichments resulting from selection, given the results of our simulations (fig. 3*A* and *C*). We thus reasoned that another aspect of the data, such as genotype calling errors, might have decreased the power to detect selection in the pilot data. We tested this hypothesis using the 1000G Phase 1 data set, in which genotype quality and coverage were improved for African (AFR) and Asian (ASN) samples (Abecasis et al. 2012). Using this data set, we retrieved a signal in the Asian sample, which displayed highly significant enrichment ($OR_C = 1.49$, table 1). This finding clearly indicates that DIND is insensitive to low coverage (e.g., 4.3× for the Phase 1 ASN sample) but highlights the ability of the genotype calling errors inherent to low-coverage data sets to wipe out the selection signal.

By contrast, no significant enrichment of genic SNPs was observed among iHS outliers in any of the various data sets studied (table 1). This finding contrasts with previous results for iHS and the HapMap genotyping data set, reporting highly significant enrichments (Voight et al. 2006). We first investigated whether our methodology (i.e., resampling scheme for significance calculation and window size) could account for such an absence of enrichment (supplementary text, Supplementary Material online). We found that the enrichment was replicated when our method was applied to the HapMap data and that the results obtained were independent of the window size used (supplementary fig. S6, tables S8 and S9, Supplementary Material online). Our results may, therefore, simply reflect an inadequacy of iHS to detect an enrichment in putative targets of selection in the specific context of the WGS data set used (i.e., low coverage and poor genotype calling quality, or small sample size). When we used the 1000G Phase 1 data set, for which the genotype calling quality was higher, we observed a slightly significant enrichment in Europeans, despite the similar mean coverage between the two 1000G data sets (4.5× for Phase 1 vs. 5.1× for Pilot) (table 1). These results suggest that iHS, which we

Table 1. Enrichment of Genic and Nonsynonymous SNPs, as Opposed to Nongenic SNPs, among iHS and DIND Outliers Calculated on 100-kb Windows.

Population	OR	Genic vs. Nongenic						Nonsynonymous vs. Nongenic					
		1000G Pilot		CG		1000G Phase1		1000G Pilot		CG		1000G Phase 1	
		DIND	iHS	DIND	iHS	DIND	iHS	DIND	iHS	DIND	iHS	DIND	iHS
AFR	OR	1.34***	0.76	1.27***	0.96	1.50***	0.87	1.23**	0.68	1.40***	0.98	1.53***	0.68
	OR_C	1.27***	0.81	1.09**	0.98	1.28***	0.92	1.18	0.72	1.17*	1.00	1.47***	0.67
EUR	OR	1.22***	0.82	1.13***	1.01	1.34***	0.96	1.25**	1.02	1.04	1.04	1.30**	1.10
	OR_C	1.22***	1.02	1.28***	1.07	1.29***	1.11*	1.22*	1.21	1.13	1.09	1.24**	1.17
ASN	OR	0.95	0.79	0.86	1.02	1.38***	0.90	1.01	0.63	0.95	0.65	1.33***	0.91
	OR_C	0.97	0.96	0.98	1.07	1.49***	1.02	1.16*	0.86	1.05	0.69	1.62***	0.98

NOTE.—OR_C indicates that the logistic regression used to calculate the OR controlled for the following confounding factors: mean recombination rate, mean coverage, and number of SNPs per window.

*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

found to be slightly more sensitive to coverage than DIND (figs. 2 and 3), is also sensitive to the genotype calling quality of the data.

When several potentially confounding factors were taken into account, our analyses showed that DIND was more powerful than iHS for detecting an enrichment of certain SNP functional classes among putative targets of selection, in the context of both small sample sizes (CG data set) and low genotype calling quality/low coverage (1000G Pilot and Phase 1 data sets). Given that our simulations showed that DIND was mostly insensitive to other selective regimes, these enrichments probably reflect the action of recent, strong positive selection.

Genomic Regions Enriched in Selection Signals Present Robust Evidence of Positive Selection

To assess whether genomic regions enriched in selection signals are biologically meaningful, we considered the distribution of SNPs presenting outlier iHS and DIND values along the genome, to detect regions potentially targeted by selection, given that positive selection tends to create local clustering of outliers (Voight et al. 2006). We evaluated the extent to which these regions overlap with genes displaying identified, robust signals of positive selection. Specifically, we searched for regions presenting the greatest clustering of extreme values of iHS and DIND (i.e., the 1% of 100-kb sliding windows presenting the highest proportion of iHS or DIND outliers) in each data set, 1000G Pilot, 1000G Phase 1, and CG (supplementary tables S10–S12, Supplementary Material online). Because the number of SNPs varies across sliding windows, we grouped windows into bins presenting similar numbers of SNPs and determined the 1% highest proportion of iHS or DIND outliers separately for each bin (Voight et al. 2006). As iHS and DIND were found to have maximum power for the detection of positive selection with high $2Ns$, our set of genes with extreme outlier clustering should overlap, to a large extent, the regions of the genome previously found to present robust signatures of strong positive selection. We thus compared it to 1) the list of genes presenting signals consistent with the hard sweep model ($2Ns = 200–800$) using iHS (Voight et al. 2006) and 2) the top list of candidate genes obtained with various linkage disequilibrium (LD)-based statistics, for example, LRH and XP-EHH (Sabeti et al. 2007). The genes with extreme DIND outlier clustering in the 1000G Pilot, 1000G Phase 1, and CG data sets included 42%, 50%, and 42%, respectively, of the top genes detected by these previous studies, whereas iHS detected only 19%, 34%, and 19%, respectively (table 2).

DIND retrieved well-known signals of positive selection, some of which were strongly supported by functional data (fig. 4*A* and *B*; supplementary figs. S7–S11, Supplementary Material online). For example, DIND consistently identified, across the three WGS data sets, the emblematic case of the rs4988235 mutation in the lactase (*LCT*) gene region, which is known to be associated with persistence of lactase activity in adulthood (Enattah et al. 2002; Bersaglieri et al. 2004; Kelley and Swanson 2008). This mutation is the core SNP of a 100-kb window located in the peak of the DIND signal and containing the second highest proportion of SNP outliers of the *LCT* region (fig. 4*A*; supplementary figs. S7–S11, Supplementary Material online). Likewise, DIND retrieved the well-known cases of the *ADH* cluster and the *EDAR* gene (table 2) (Osier et al. 2002; Carlson et al. 2005; Sabeti et al. 2007; Barreiro et al. 2008). For *EDAR*, the signal retrieved from the 1000G data set encompassed the nonsynonymous V370A mutation (rs3827760), which has been associated with hair thickness, tooth morphology, and the number of eccrine sweat glands (Fujimoto et al. 2008; Kamberov et al. 2013) (fig. 4*B*; supplementary figs. S10–S11, Supplementary Material online). Conversely, none of these emblematic cases of selection was detected by iHS in the CG and 1000G Pilot data sets (table 2, fig. 4*A* and *B*; supplementary figs. S7–S9, Supplementary Material online), with the exception of the *LCT* region in the CG data, but only if 1-Mb windows were used (supplementary fig. S9*A*, Supplementary Material online). The strong selection signals of both *EDAR* and *LCT* were restored when iHS was applied to the 1000G Phase 1 data set (supplementary figs. S10*B* and S11*A* and *B*, Supplementary Material online). These examples highlight again the sensitivity of iHS to both low coverage and the quality of genotyping calls. Consistent with the results of enrichment analyses among functional SNP classes, DIND did replicate signals of strong, recent positive selection more effectively than iHS.

Functional and Medical Relevance of Regions Enriched in Signals of Selection

To provide additional support to the adaptive significance of the genomic regions enriched in selection signals, we next investigated the extent to which these regions were enriched in SNPs that are likely to have functional consequences, that is SNPs associated with phenotype traits or disease by genome-wide association studies (here termed as GWAS-SNPs, see Materials and Methods) (Hindorff et al. 2009). Given the higher performance of DIND, with respect to iHS, with WGS data sets, we restricted our analysis to DIND outlier regions. We found a genome-wide enrichment of GWAS-SNPs in Africans and, even more so, in Europeans, as expected, given that most GWAS have been performed in populations of European descent (fig. 5*A* and *B*; supplementary table S13, Supplementary Material online). Likewise, when focusing on particular traits or diseases overrepresented among DIND outliers in each population, Europeans displayed the highest number of enriched categories (supplementary table S14, Supplementary Material online). For example, various outlier SNPs were found to be associated with skin pigmentation, such as rs1667394 A/G in *OCA2* or rs916977 A/G in *HERC2*, for which the selected allele is associated with fairer skin, hair, and eye color (supplementary table S15, Supplementary Material online). This observation is consistent with a Gene Ontology (GO) analysis in which subcategories relating to pigmentation are enriched in genes with extreme DIND outlier clustering in Europeans (e.g., melanocyte differentiation, pigment cell differentiation, supplementary table S16, Supplementary

Table 2. Overlap of Extreme iHS and DIND Outlier Clustering, Calculated from the 1000G Pilot and Phase 1 and CG Data Sets, with Regions Previously Found to Present Robust Signatures of Positive Selection.

Position	Population	Gene	Voight et al. (2006)[a]	Sabeti et al. (2007)[a]	1000G Pilot[b]		CG[b]		1000G Phase 1[b]	
					DIND e-value[c]	iHS e-value	DIND e-value	iHS e-value	DIND e-value	iHS e-value
1q23.3-q24	ASN	*BLZF1, SLC19A2*		LRH, iHS	1.0000	0.0262	0.0005	1.0000	0.0108	0.2367
1p31.3	ASN	*SLC44A5*	iHS		0.0074	0.0441	0.0037	0.0932	0.0153	0.0898
1p34.3	EUR	*NCDN, TEKT2*	iHS		0.1058	1.0000	1.0000	1.0000	0.0178	1.0000
2q13	ASN	*EDAR*		LRH, iHS, XP-EHH	0.0006	0.1026	0.0422	1.0000	0.0011	0.0013
2q21.3	ASN	*SULT1C* cluster	iHS		0.0019	0.0108	0.0011	0.1469	0.0006	0.0054
2q21.3-q22.1	EUR	*LCT*	iHS	LRH, iHS, XP-EHH	0.0002	0.1205	0.0003	0.0172	0.0002	0.0046
2p23.3	AFR	*NCOA1, ADCY3*	iHS		0.0203	0.0069	0.0169	0.0872	0.0054	0.0286
2q31.2	ASN	*PDE11A*		LRH, iHS, XP-EHH	0.0219	0.0714	0.0241	0.0009	0.0084	0.0291
	EUR				0.0106	1.0000	0.0031	0.1755	0.0197	0.0806
4p13	ASN	*SLC30A9*		LRH, iHS, XP-EHH	0.0052	0.0593	1.0000	1.0000	0.0000	0.0326
4q21-23	ASN	*ADH* cluster	iHS		0.0058	0.0735	0.0008	0.0859	0.0085	0.0395
8q11.21-23	AFR	*SNTG1*	iHS		0.0021	0.0340	0.0291	0.0222	0.0011	0.0621
	ASN		iHS		0.0021	0.0003	0.0011	0.0421	0.0014	0.0308
	EUR		iHS		0.0011	0.0764	0.0199	0.0307	0.0026	0.0089
9p22.3	ASN	*C9orf93*	iHS		1.0000	0.2567	1.0000	0.197	0.0572	0.0436
10q21.1	ASN	*PCDH15*		LRH, iHS, XP-EHH	0.0171	0.0004	0.0163	0.0024	0.0024	0.0021
12q21.2	AFR	*SYT1*	iHS		0.0008	0.0003	0.0036	0.0003	0.0017	0.0003
15q21.1	EUR	*SLC24A5*[d]		XP-EHH	NA	NA	NA	NA	NA	NA
15q22	ASN	*HERC1*		XP-EHH	9e-05	0.1508	1.0000	1.0000	0.0001	0.0233
16q22.3-q23.1	AFR	*CHST5, ADAT1, KARS*		LRH,iHS	0.0535	1.0000	0.0159	0.0085	0.0278	0.0742
	ASN				0.0997	1.0000	0.0284	0.0868	0.0815	0.0252
17q23	EUR	*BCAS3*		XP-EHH	0.0057	0.0421	0.0032	0.0891	0.0267	0.0007
20cen	AFR	*SPAG4*	iHS		0.0137	0.0698	0.0444	0.2307	0.0131	0.0374
	EUR		iHS		0.0144	1.0000	0.0010	0.011	0.0069	0.0129
20cen	ASN	*ITGB4BP, CEP2*	iHS		1.0000	0.0792	1.0000	1.0000	1.0000	0.0385
22q12.3	AFR	*LARGE*		LRH	1.0000	1.0000	0.1153	0.0284	0.0395	0.0059

[a]Empty cells correspond to regions not present in the list of the top selection targets in these studies.

[b]The genes with at least one window showing extreme proportion of outliers are underlined (the windows are grouped into bins with similar numbers of SNPs, and the 1% most extreme proportion of outliers are determined separately for each bin).

[c]The e-value is based on the calculation of the proportion of outliers within sliding windows of 100 kb, centered on each SNP (outlier clustering). The e-value is the genome-wide proportion of windows, with an outlier clustering greater than the maximum clustering value observed for the gene.

[d]For *SLC24A5*, NA indicates that no SNP with a DAF over 0.2 was found in this gene. Note that the contiguous genes *SLC12A1* and *FBN1*, which are located 80 kb and 250 kb away from *SLC24A5*, respectively, were detected using the 1000G Pilot and Phase 1 data sets.

Material online). Similarly, seven SNPs associated with height were among those presenting the strongest signals of positive selection in Europeans, five of which have been associated with increased height. Likewise, four SNPs associated with height were found to be under positive selection among Africans and four among Asians (supplementary table S17, Supplementary Material online). Finally, one SNP associated with age at menarche was among the strongest signals of positive selection in Europeans and three in Africans, all the selected alleles being associated with an older age at menarche onset (supplementary table S18, Supplementary Material online).

For disease-associated SNPs, several categories, such as immune-related diseases and cancers, were overrepresented in DIND outliers. Interestingly, for some of the GWAS-SNPs associated with immune-related disorders, we observed a clear directionality (e.g., risk or protection) of the selective pressure, with most of the selected alleles increasing disease risk (~70%, table 3; supplementary table S19, Supplementary Material online). By contrast, no clear directionality of the selection pressure was observed for GWAS-SNPs associated with other human diseases, including cancers (table 3; supplementary table S19, Supplementary Material online) or diet-related traits, such as fat metabolism and cholesterol levels (supplementary table S20, Supplementary Material online).

Discussion

The aim of this study was not to perform a hypothesis-generating genome-wide scan of selection using classical outlier approaches. The overlap of outlier loci among existing studies remains limited (Akey 2009), owing to the heterogeneity of statistics used, threshold definitions of "outlier," time frames of the selective events recovered, and high false discovery rates (FDRs) (Kelley et al. 2006; Teshima et al. 2006), emphasizing the need for studies that consider demography and other selective models. Our aim here was instead to provide a global assessment of the genome-wide prevalence of recent,

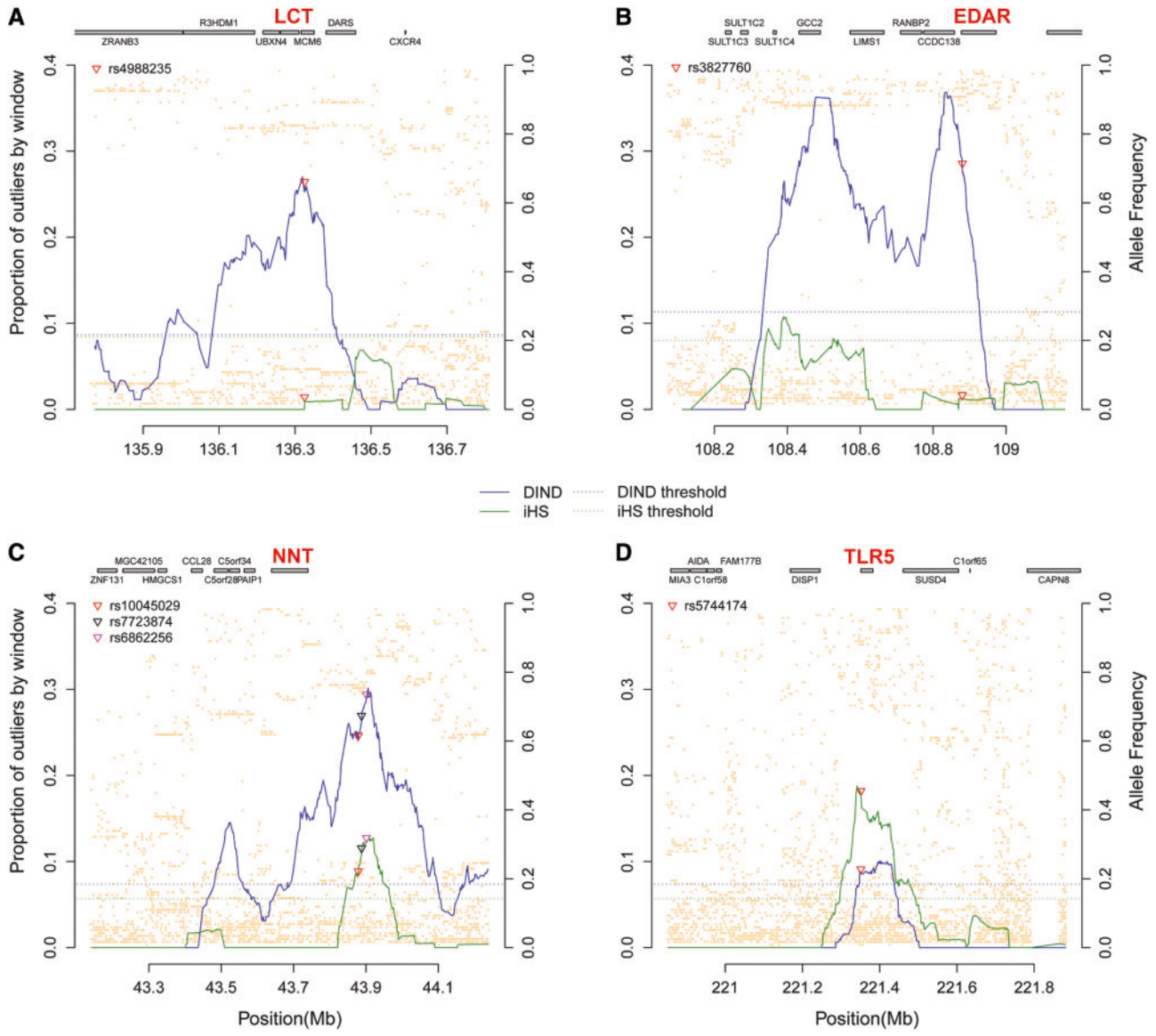


Fig. 4. Examples of candidate genomic regions under positive selection in the 1000G Pilot data set. iHS and DIND were calculated for 100-kb windows. Lines show the proportion of iHS (in green) and DIND (in blue) outliers by window. The dotted lines represent, for iHS and DIND, the threshold defining the 1% most extreme proportions of outliers by window (100 kb). The orange dots are the DAFs. The gray rectangles show the position of the genes. (*A*) *LCT*. Evidence of positive selection in the EUR population at locus 2q21, centered on SNP rs4988235, responsible for lactase persistence in adulthood (red triangle). (*B*) *EDAR*. Evidence of positive selection in the ASN population at locus 2q13, around the SNP rs3827760, associated with hair morphology (red triangle). (*C*) *NNT*. Evidence of positive selection in the AFR population at locus 5p12, implicated in familial glucocorticoid and cortisol deficiency, and particularly around the SNPs rs10045029, rs7723874, and rs6862256, associated with *NNT* expression (red, dark blue, and magenta triangles, respectively). (*D*) *TLR5* region. Evidence of positive selection in the AFR population at locus 10q24, involved in the recognition of bacterial flagellin, and, in particular, around SNP rs5744174, a nonsynonymous mutation (L616F) associated with lower levels of NF-kB signaling in response to flagellin (red triangle).

strong positive selection as a mechanism of adaptive change in humans. We found that the haplotype-based iHS and DIND statistics are both powerful to detect hard sweeps, in the context of WGS data sets and human demography, and insensitive to background selection and other modes of selection. By applying these statistics to WGS data sets, we provide evidence of positive selection targeting specific functional SNP classes, that is, enrichments of genic and nonsynonymous SNPs among selection signals, and that such selection signals are enriched in SNPs associated with phenotypic variation.

First, our simulation study showed that the haplotype-based statistics iHS and DIND are powerful to detect selection over a large range of allele frequencies. We also found that the power of these statistics remained virtually unchanged when simulating variation of mutation and recombination rate, and that it is not affected by the reconstruction of gametic phases. Notably, our simulation study demonstrated the almost total

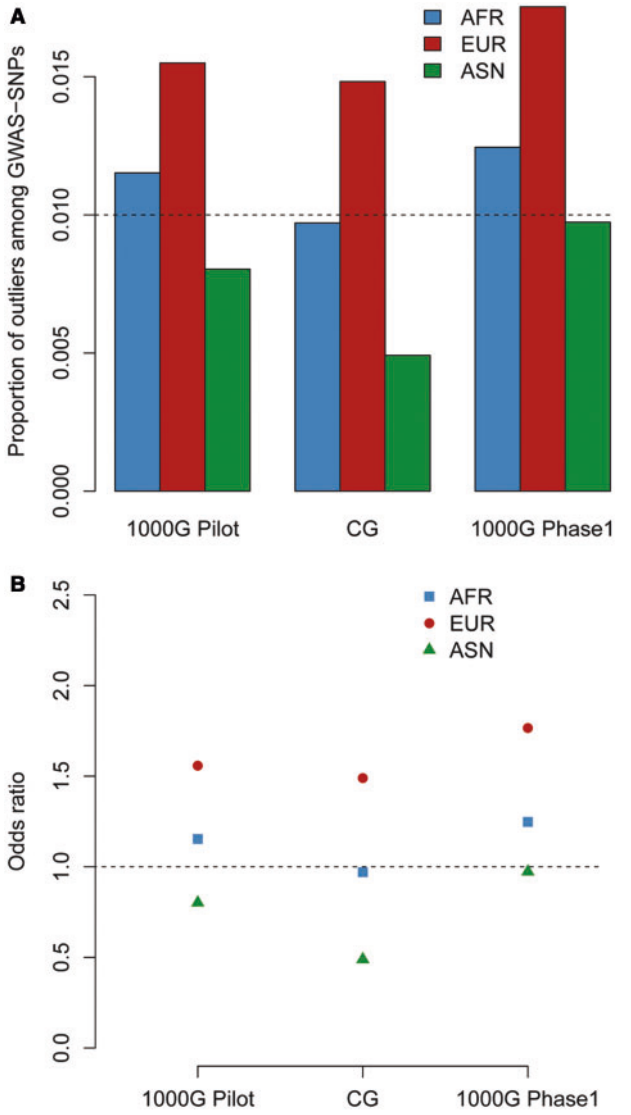


FIG. 5. Enrichment in GWAS-SNPs among DIND outliers. DIND was calculated for 100-kb windows (results for DIND calculated for 1-Mb windows are available in supplementary table S13, Supplementary Material online). GWAS-SNPs were filtered for *P* value lower than 10^{-7}. A single entry was retained for each SNP-trait association, and LD was accounted for (see Materials and Methods). (*A*) Proportion of GWAS-SNPs that are DIND outliers. Bar plots show the proportion of outliers among GWAS-SNPs for each data set and each population (from left to right, AFR, EUR, ASN). The black dotted line indicates the proportion of outliers among all the SNPs of the genome. (*B*) Enrichments of GWAS-SNPs among DIND outliers. Relative enrichment of GWAS-SNPs among DIND was measured using OR. The black dotted line corresponds to an OR equal to 1. An OR equal to or smaller than 1 indicates no enrichment. An OR greater than 1 indicates enrichment of GWAS-SNPs among DIND outliers (EUR in all data sets, and AFR in 1000G Pilot and Phase 1 data sets).

insensitivity of DIND to low coverage (as low as 3×). Indeed, because the nucleotide diversity π is not particularly sensitive to low-frequency variants DIND is particularly insensitive to low coverage mainly affecting these low-frequency variants. Furthermore, we showed that the power to detect hard sweeps was greatest for the 1000G data set, because sample

Table 3. Numbers of Selected Risk Alleles, Protection Alleles, and Nonreported Effect Alleles for Diseases for Which DIND Outliers Displayed Enrichment.

Disease categories	Population		Risk[a]	Protection[b]	NR[c]
Immune-related diseases	ALL	Count[d]	13	6	0
		Percent	68.42%	31.58%	0.00%
	AFR	Count[d]	1	2	0
		Percent	33.33%	66.67%	0.00%
	EUR	Count[d]	9	3	0
		Percent	75.00%	25.00%	0.00%
	ASN	Count[d]	3	1	0
		Percent	75.00%	25.00%	0.00%
Cancer	ALL	Count[d]	7	7	1
		Percent	46.67%	46.67%	6.67%
	AFR	Count[d]	0	4	0
		Percent	0.00%	100.00%	0.00%
	EUR	Count[d]	3	2	1
		Percent	50.00%	33.33%	16.67%
	ASN	Count[d]	4	1	0
		Percent	80.00%	20.00%	0.00%
Other diseases	ALL	Count[d]	6	8	6
		Percent	30.00%	40.00%	30.00%
	AFR	Count[d]	2	2	1
		Percent	40.00%	40.00%	20.00%
	EUR	Count[d]	2	4	3
		Percent	22.22%	44.44%	33.33%
	ASN	Count[d]	2	2	2
		Percent	33.33%	33.33%	33.33%

[a]The selected allele (derived allele) is associated with a higher risk of developing disease.
[b]The selected allele (derived allele) is not the risk allele defined in the NHGRI GWAS database.
[c]The risk allele was not reported in the NHGRI GWAS database.
[d]Counts were obtained taking LD into account.

size appeared to have a stronger effect on the power of iHS and DIND than coverage variation. Indeed, the small sample size of the CG data set was not compensated by its deep coverage (~40×) for the detection of signals of strong ongoing selective sweeps.

Second, echoing the simulation results, the analysis of the WGS data sets showed that DIND performed better than iHS in the context of small sample sizes, as shown for the CG data set, and low coverage, as shown for the 1000G data set. In this context, iHS failed to replicate the enrichment of genic SNPs among outliers previously obtained with HapMap ((Voight et al. 2006) and this study) and to detect well-known signals of positive selection (Voight et al. 2006; Sabeti et al. 2007; Pickrell et al. 2009). iHS is sensitive not only to low coverage but also to genotype calling errors. Following the improvement of data quality in the 1000G Phase 1 release, the OR_C of iHS increased and reached significance in Europeans (table 1), and some of the strongest signals of selection, including *LCT* and *EDAR*, were restored. In addition, the iHS signal-to-noise ratio is lower in WGS data sets than in genotyping data sets, because extended haplotypes are more rapidly broken in the presence of low-frequency variants (Grossman et al. 2013). That low-frequency variants are more common in WGS (1000 Genomes Project Consortium 2010) could explain the absence (or weakness) of enrichment in genic SNPs within iHS outliers, while such enrichment has been observed in

genome-wide SNP data sets. This would not affect DIND, as π is not particularly sensitive to low-frequency variants, consistent with significant enrichments of genic SNPs among DIND outliers only. Likewise, when simulating DNA sequence data under selection, we observed a lower clustering of outliers for iHS with respect to DIND. In addition, the breakdown of extended haplotypes by low-frequency variants is exacerbated in genic regions, where a higher proportion of low-frequency variants is observed (Abecasis et al. 2012), because moderately deleterious mutations are maintained at low frequency by negative selection (30–42% of human nonsynonymous mutations are moderately deleterious, $0.01\% < |s| < 1\%$, see (Boyko et al. 2008)). Accordingly, we obtained lower OR_C for Africans than for non-Africans with iHS, for all data sets, as expected, given that negative selection is more efficient in populations with large effective sizes.

The enrichments of genic and nonsynonymous SNPs among DIND outliers, and its substantial power to detect well-known signals of selection, provide an important proof-of-concept of the detection of genuine positive selection events in WGS data sets. For example, we identified the functionally validated selection signal at *TLR5* in Africans (Grossman et al. 2013) (fig. 4*D*; supplementary figs. S7–S11, Supplementary Material online). *TLR5* is an innate immunity receptor involved in the recognition of bacterial flagellin, highlighting the importance of the selective pressures imposed by pathogens during human evolution (Barreiro and Quintana-Murci 2010; Quintana-Murci and Clark 2013). In addition, several new signals are of particular interest because they involved SNPs predicted to be functional by other studies (supplementary tables S10–S12, Supplementary Material online). For example, we detected a strong signal on chromosome 5p12 in the African population, for all WGS data sets (fig. 4*C*; supplementary figs. S7–S11, Supplementary Material online). The peak signal was located 150 kb downstream from the *NNT* gene, and the SNPs with the highest DIND scores have been associated with *NNT* expression (i.e., expression quantitative trait loci, eQTLs) in Africans ($P = 6.4 \times 10^{-8}$; [Pickrell et al. 2010]). *NNT* has recently been implicated in familial glucocorticoid deficiency, which triggers low cortisol levels, hypoglycemia, and hyperpigmentation (Meimaridou et al. 2012). Glucocorticoids are steroid hormones that mediate homeostatic responses to environmental stressors, and these responses are known to vary among human populations (Maranville et al. 2011). Finally, among the strongest selection signals in east Asians, three gene regions have been linked to breast cancer. These include *RAD51L1* and the *ECHDC1-RNF146* region identified by GWAS (Hoggart et al. 2007; Gold et al. 2008), and *HERC1*, which has previously been reported as a selection target and is mutated in breast cancer (Grossman et al. 2013). These observations highlight the need for further studies to better understand the extent to which cancer, which is generally a rather late-onset disease, has been a selective factor by itself or a by-product of other selective forces exerting pressure on pleiotropic genes.

Further support to the adaptive significance of the genomic regions enriched in selective signals came from the overlap with GWAS, providing new insight into the relationship between past selection and benign and disease-related phenotypic variation. We found global enrichments in GWAS-SNPs among DIND outliers, supporting again the notion that we may detect true selective events from WGS data. Importantly, we were able to infer the phenotypic directionality of selective events in some cases. For example, although it has been suggested that height-associated SNPs are subject to polygenic adaptation by weak selection (Turchin et al. 2012), we detected five SNPs associated with this polygenic trait that displayed signatures of strong selection favoring high stature in European populations (supplementary table S17, Supplementary Material online). Likewise, we detected four SNPs in African and European samples for which positive selection has favored a later onset of menarche (supplementary table S18, Supplementary Material online). It has been suggested that the increasing complexity of human societies (e.g., the emergence of farming) has delayed psychosocial maturity (Gluckman and Hanson 2006) and that the occurrence of sexual maturity in psychosocially immature females is detrimental. Our analyses suggest that selection has acted to compensate for this trend by shifting sexual maturity to older ages. Importantly, we also observed a strong skew in selection, targeting alleles associated with a higher risk of immune-related diseases. Our results further support the hypothesis that the incidence of immune-related disorders in modern societies may at least partly reflect the consequences of past selection for stronger immune responses to combat infection (Barreiro and Quintana-Murci 2010; Raj et al. 2013).

More generally, our results must be seen in the context of recent debates as to the prevalence of hard sweeps in the human genome. Two recent studies have suggested that classic selective sweeps have been relatively rare during human evolution (Hernandez et al. 2011; Granka et al. 2012) and that most of these "sweeps" could be explained by the widespread action of background selection (Hernandez et al. 2011). Here we show that the two haplotype-based statistics used are robust to background selection and underpowered for the detection of positive selection events other than hard, or nearly hard, sweeps. Our results should, therefore, highlight only the occurrence of recent, strong positive selection. However, although clearly significant, the ORs obtained (OR_C) in the enrichment analyses for functional SNP classes were generally modest (1.2–1.5), supporting the notion that the prevalence of such sweeps is moderate. For example, an OR_C of 1.25 indicates that no more than 20% of candidate genes are true targets of positive selection. This observation may explain the limited overlap of outlier loci among other studies (Akey 2009), as well as between DIND and iHS in this study. However, using the OR_C of the 1000G Phase 1 data set, we can roughly estimate the number of genes under selection at approximately 70–100 in each of the different population groups. However, these numbers may represent the lower bound of genes under selection, given that the actual power to detect selection is lower than 100% and that we neglect the occurrence of selection in nongenic regions, for example, overlap of iHS and eQTL signals (Kudaravalli et al. 2009). Taken together, our results indicate that recent, hard

sweeps have played a moderate, but significant, role over the last ~60,000 years of human evolution. Given that positive selection regimes other than the hard sweep model, such as polygenic adaptation by weak selection and selection on standing variation, cannot be detected by our approach, the degree of positive selection *lato sensu* acting on the human genome is undoubtedly higher than suggested here and in previous studies.

We conclude that low-coverage WGS data can be efficiently used for the detection of selective sweeps, revealing genes and functions accounting for adaptive phenotypic variation in humans or other species. The development of methods that can be safely used in the context of low-coverage data is of particular importance for the design of population genetic studies, as the sequencing of many individuals at high coverage remains costly. It is now time to refine analyses by focusing on populations living in extreme environmental conditions—high altitude, Artic climate, forest- or savannah-based populations—or with different modes of subsistence. Whole-genome sequencing of individuals from these populations, even at low coverage, should improve our understanding of the genetic basis of human adaptation to specific environments.

Materials and Methods

Data

The low-coverage part of the 1000G Pilot Project consists of data for samples from four populations: 59 unrelated Yoruba from Ibadan, Nigeria (AFR), 60 unrelated Utah residents with Western and Northern European ancestry (EUR), and 60 unrelated Asians (ASN), 30 Han Chinese from Beijing and 30 Japanese from Tokyo. All these samples were sequenced at low mean coverage: 3.7× for the AFR panel, 5.1× for EUR panel, and 2.8× for the ASN panel. In total, we analyzed 9,760,562 SNPs for AFR, 6,858,242 SNPs for EUR, and 5,674,252 SNPs for ASN. The ancestral state of each SNP was retrieved from the 1000G Project website (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/release/2010_03/pilot1, last accessed April 9, 2014).

We evaluated the influence of genotype call quality by incorporating into our analyses a subset of the Phase 1 data of the 1000G Project, in which the quality of genotype calls was significantly improved. For comparison purposes, we included only individuals for whom data were already present in the Pilot release. Our subset of the 1000G Phase 1 data set consisted of 52 AFR, 45 EUR, and 58 ASN individuals (supplementary table S6, Supplementary Material online). These samples were sequenced at low mean coverage: 4.4× for the AFR panel, 4.5× for EUR panel, and 4.3× for the ASN panel. In total, we analyzed 12,848,493 SNPs for AFR, 7,577,087 SNPs for EUR, and 7,161,377 SNPs for ASN. We rendered the Pilot and Phase 1 data sets comparable, by removing the singletons from the 1000G Phase 1 data set (supplementary fig. S5, Supplementary Material online). The ancestral state of each SNP was retrieved from the 1000G Project website (ftp://ftp.ncbi.nih.gov/1000genomes/ftp/technical/working/20120316_phase1_integrated_release_version2/, last accessed April 9, 2014).

We also studied the high-coverage data of the CG public data set (software version 1.10.0.26). We selected samples from nonadmixed populations only and pooled together populations presenting close genetic affinities, to increase sample size (supplementary table S6, Supplementary Material online). We pooled together nine Yoruba from Ibadan, Nigeria, and four Luhya from Webuye, Kenya, to form a single panel of 13 unrelated Africans (AFR). We pooled together nine Utah residents with northern and western European ancestry from the centre d'etude du polymorphisme humain (CEPH) collection and four individuals from Tuscany, Italy, to form a single panel of 13 unrelated Europeans (EUR). We pooled together four Han Chinese from Beijing, China, and four Japanese from Tokyo, Japan, to form a single panel of eight unrelated Asians (ASN). All these samples were sequenced with a high mean coverage of over 50×. We removed from the analysis all SNPs presenting 5% or more low-quality Illumina calls (i.e., calls with a mapping and assembly with qualities [MAQ] mapping quality of 0). We also removed from the analysis the 11q region in which we found an accumulation of Mendelian errors. In total, we analyzed 10,070,271 SNPs for AFR, 6,281,785 SNPs for EUR, and 5,065,417 SNPs for ASN. The ancestral states of the SNPs were determined from the ancestral sequence provided by the 1000G Project and the genomes of five primates: gorilla, chimpanzee, orangutan, macaque, and marmoset (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/, last accessed April 9, 2014).

A Realistic Human Demographic History

We aimed to simulate data under a realistic demographic scenario that was tractable with the forward-in-time simulation software SFS-CODE (Hernandez 2008). We determined realistic values of key demographic parameters by comparing observed data (i.e., 20 independent 1.33-kb noncoding regions previously resequenced in 95 Africans, 60 Europeans, and 60 Asians, see Laval et al. [2010]), with simulations of various demographic scenarios obtained with SFS-CODE, summarizing each data set in terms of the mean and standard deviation of several statistics (Tajima's D, the number of segregating sites S and F_{ST}) (supplementary text and table S1, Supplementary Material online). We simulated a 1.33-kb DNA fragment with θ ($\theta = 4N\mu$, μ is the per generation per site mutation rate) and ρ ($\rho = 4Nr$, r is the per generation rate of recombination between adjacent loci) equal to 0.001. We simulated three populations mimicking the African, European and Asian populations and tested several scenarios by varying the age and strength of bottlenecks and expansions. For each simulation, 95 individuals were sampled from the African population, and 60 were sampled from each non-African population (10,000 simulations for each scenario). We used an ABC approach (Beaumont et al. 2002) to estimate the posterior probability of each demographic model, as previously described (Laval et al. 2010) (supplementary text, Supplementary Material online). We retained the

demographic scenario with the highest posterior probability in order to perform all subsequent simulations, that is, recent selective sweep, background selection, interaction of recent selective sweep, and background selection as well as the neutral simulations that were used to determine the thresholds applied to detect selection.

Consistent with the general model of human evolution (Voight et al. 2005; Laval et al. 2010; Gravel et al. 2011), the retained scenario consisted of an ancestral African population of constant size ($N = 10{,}000$) that split into two populations (African and non-African) 60,000 years ago (fig. 1*A*). An expansion resulted in an instantaneous 50 times increase in the African population, 20,000 years ago. This time frame corresponds to the mean of the times corresponding to the Bantu expansions (Diamond and Bellwood 2003) and a more ancient expansion that may have occurred in Africa (e.g., ~30,000 years ago [Voight et al. 2005; Laval et al. 2010]). The bottleneck accompanying the out-of-Africa exodus caused an instantaneous decrease in the ancestral non-African population, which was halved. This population then split again into two populations (European and Asian) 20,000 years ago. Finally, both these populations underwent an instantaneous 100-fold expansion 6,000 years ago, corresponding to the Neolithic expansion (Laval et al. 2010). The migration rate (m) was set to 1.3×10^{-5} and was fixed according to what is commonly admitted concerning modern human evolution. We minimized computation time by using an ancestral effective size $N = 100$, although the effective population size for humans is generally considered to be $N = 10{,}000$. Indeed, if it is desired to simulate over t generations a population with parameter values N, μ, ρ, and s, then a simulation using instead N/λ, $\lambda\mu$, $\lambda\rho$, and λs, evolved for t/λ generations, for some $\lambda > 1$, will generate approximately the desired AFS and patterns of LD (Hoggart et al. 2007). Consequently, the AFS simulated using SFS-CODE are not affected by the simulated population size (Hernandez 2008) (see also the SFS-CODE documentation). In addition, we tested the effect of this scaling on the power of statistics based on the levels of LD surrounding a positively selected allele such as iHS and found no effect (data not shown).

Simulating Full Sequence Data

We used SFS-CODE to simulate DNA regions according to the demographic model, mutation, and recombination rates described above. We used this calibrated demographic model to perform all subsequent simulations, that is, all neutral simulations as well as those under the various models of selection investigated. For each simulation, 59 individuals were sampled from the African population and 60 from one of the two non-African populations, for matching with the 1000G Pilot samples (largest number of sampled individuals for the data sets analyzed). We first simulated neutrally evolving DNA regions and positive selection models, assuming the hard sweep model (Pritchard et al. 2010). A new advantageous mutation with a population genetics selection parameter $2Ns$ was inserted into the middle of the sequence, at a frequency of $1/2N$, in a specific population, at a specific time t. We simulated 100-kb DNA regions with $2Ns$ equal to 100, a combination of parameters, that is, length of DNA region and strength of selection, which was previously used to consistently estimate the power to detect recent positive selection in humans (Voight et al. 2006; Barreiro et al. 2009). The time t was drawn from a range of recent values (7,500; 10,000; 15,000; 20,000; 25,000; 37,500; 50,000) to obtain a large range of SAF ($0 \leq \text{SAF} \leq 1$) values, covering the frequency spectrum from 0 to 1.

We then simulated background selection. We assumed that 20% of the mutations of each 100-kb region were negatively selected, and we explored a wide range of $2Ns$ ranging from -500 to -1. We also simulated models of interaction between positive and background selection. To do so, a new advantageous mutation ($2Ns = 100$) was inserted (frequency of $1/2N$), in a specific population at a specific time t (same range of recent values as above), into the middle of a sequence, where 20% of sites were set as negatively selected with identical $2Ns$ values. We explored various $2Ns$ values including $2Ns = -1$, $2Ns = -100$, and $2Ns = -500$.

We also aimed to simulate scenarios of positive selection on standing variation. Unfortunately, to our knowledge, it is not possible to simulate positive selection on standing variation with SFS-CODE. We therefore used mpop (Pickrell et al. 2009) for forward simulations, assuming a population of constant effective size ($2N = 1{,}000$ chromosomes). Indeed, mpop can simulate positive selection on standing variation only for populations of constant size. We set the per locus mutation rate ($\theta = 4N\mu$) and the rate of recombination between adjacent loci ($\rho = 4Nr$) to 0.001, as in previous studies. We simulated standing variation scenarios by adding a selective advantage of $s = 0.1$ ($2Ns = 100$) to a previously neutral allele of frequency 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, or 0.5.

Simulating Next-Generation Sequencing Data

To simulate low-coverage data, we used the short-read simulator ShotGun (Kang et al. 2013). It allowed us to simulate 100-bp reads, with realistic read depth distribution following a negative binomial distribution, which is a Gamma mixture of Poisson. Indeed, the read depth distribution is known to follow a Poisson distribution, but stochastic and experimental limitations result in overdispersed read depths across bases. The level of overdispersion is controlled by a shape parameter set to be equal to 4 (Kang et al. 2013). The sequencing error rate specified was set to 0.001 (Shendure and Ji 2008). In order to simulate the SNP calling step, we used Thunder (Li et al. 2011), which takes into account the LD information to call genotypes. Thunder is an extension of MaCH, the genotype imputation and phase reconstruction software (Li et al. 2010). We simulated an average coverage of $4\times$ for the African individuals, $5\times$ for the European individuals, and $3\times$ for the Asian individuals by using negative binomial distributions with means of 4, 5, and 3. These values correspond to the per individual average coverage calculated for the AFR, EUR, and ASN samples of the 1000G Pilot data set. The lower and upper bounds of the 99% confidence intervals are equal to 0–14 in Africa, 1–17 in Europe, and 0–11 in Asia. After simulating

coverage and SNP calling steps, we then removed every singleton, as observed in the 1000G data set. In addition, for each of these simulated data sets, we reconstructed the gametic phase of each individual using Thunder/MaCH and SHAPEIT (Delaneau et al. 2012), without the use of genealogical information.

We simulated small sample sizes by randomly drawing individuals from the simulations under both neutrality and positive selection described above. We randomly drew 13 individuals from the African and one of the non-African populations and eight from the other non-African population, for matching to CG data.

Statistics

To detect mutations targeted by recent positive selection, we used the haplotype-based statistics iHS and DIND (Voight et al. 2006; Barreiro et al. 2009), which are population- and SNP-specific. They were designed to directly detect mutations targeted by recent positive selection, in contrast with other approaches (e.g., AFS-based statistics, such as Tajima's *D*) that cannot identify the local target of selection because they are calculated over a given region. In addition, iHS was designed to determine whether the ancestral or derived allelic state of each mutation has been targeted by recent positive selection, whereas DIND detects positive selection on the derived allele only. Both methods are based on the same principle: the comparison of haplotypes carrying the ancestral allele with haplotypes carrying the derived allele of a given SNP.

The iHS statistic is therefore calculated when the ancestral and derived allelic states are known unambiguously. This statistic is based on extended haplotype homozygosity (EHH) (Sabeti et al. 2002), a statistic assessing the identity of haplotypes carrying the ancestral or derived alleles of a given SNP over increasing distances. It is based on the rationale that an allele targeted by strong positive selection increases in frequency much more rapidly than a neutral allele, therefore, displaying high levels of haplotype homozygosity over much greater distances than would be expected under neutrality (indeed, the neighboring region accumulates much less recombination). More specifically, the iHS is based on the integral of the observed decay of EHH (summed in both directions away from the core SNP until EHH reaches 0.05) denoted iHH. The iHS statistic is calculated as follows:

$$iHS = \frac{\ln\left(\frac{iHH_A}{iHH_D}\right) - E_p\left(\ln\left(\frac{iHH_A}{iHH_D}\right)\right)}{SD_p\left(\ln\left(\frac{iHH_A}{iHH_D}\right)\right)}$$

with iHH_A and iHH_D being the iHH calculated with haplotypes carrying the ancestral and derived alleles, respectively; E_p and SD_p are the expectation and standard deviation estimated from the empirical distribution for SNPs with a DAF p matching the frequency at the core SNP. Consequently, an extremely negative value of iHS denotes positive selection on the derived allele ($iHH_D > iHH_A$), whereas a highly positive iHS indicates positive selection on the ancestral allele ($iHH_A > iHH_D$).

The DIND is also calculated for unambiguously known ancestral and derived allelic states. This statistic is based on nucleotide diversity (π), which is used to measure the genetic diversity of haplotypes carrying the ancestral or derived allele of a given SNP. It is based on the rationale that alleles targeted by strong positive selection increase in frequency more rapidly than neutral alleles and therefore tend to have a lower nucleotide diversity than would be expected under a hypothesis of neutrality (indeed, the neighboring region accumulates fewer mutations). More specifically, DIND is the ratio π_A/π_D, with π_A and π_D being the haplotype diversity calculated with haplotypes carrying the ancestral and derived alleles, respectively. The DIND statistic is calculated as follows:

$$DIND = \frac{\pi_A}{\pi_D} = \frac{\frac{\sum_{i=1}^{n_A-1}\sum_{j=i+1}^{n_A} d_{ij}}{C_{n_A}^2}}{\frac{\sum_{k=1}^{n_D-1}\sum_{l=k+1}^{n_D} d_{kl}}{C_{n_D}^2}}$$

with n_A and n_D being the number of ancestral and derived alleles, respectively, d_{ij} being the number of differences between two haplotypes *i* and *j* carrying the ancestral allele, and d_{kl} being the number of differences between two haplotypes *k* and *l* carrying the derived allele. Consequently, very high values of DIND indicate the occurrence of positive selection on the derived allele: that is, $\pi_D << \pi_A$. Note that DIND was initially designed to capture selection targeting the derived allele but can easily be extended to detect positive selection targeting ancestral alleles. These two statistics require individual gametic phases (the effect on the power of these statistics of the phasing procedure used was evaluated, see Results).

Phasing the Data

As described above, the iHS and DIND statistics are based on haplotypic information and must therefore be calculated for individual gametic phases. For the low-coverage part of the 1000G data set (Pilot and Phase 1 releases), phased data were obtained from the MaCH website (Center for Statistical Genetics, University of Michigan, http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G-2010-06.html (last accessed April 9, 2014) for the Pilot, and http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G.2012-02-14.html (last accessed April 9, 2014) for the Phase 1 release). The phasing procedure imputed all missing genotypes, which were found at SNPs presenting 20% or more low-quality Illumina calls (i.e., calls with MAQ mapping quality of 0). For CG public data, the phased data were inferred with SHAPEIT, by merging populations (Delaneau et al. 2012). The phasing process was improved by the use of the Yoruba from Ibadan, Nigeria (YRI) trio to phase the AFR, and of 13 members of the same family (pedigree) to phase the EUR. Only the founders of AFR and EUR families were retained for positive selection analyses.

Power of iHS and DIND

The power was evaluated on 100-kb regions by simulations, assuming the demographic model described above. For each statistics, critical values were determined separately for each

population, by neutral simulations (no selected site included in the 100-kb regions), to obtain an FPR of 1%. We calculated iHS and DIND for each mutation of the 100-kb region. As the variance and mean of the DIND statistic depend on the DAF, the values of DIND can only be compared for SNPs with similar DAF values. We therefore determined the extreme values of DIND from bins of DAFs. We grouped mutations by DAF bin (from 0 to 1, in increments of 0.025) and extracted the top 1% of DIND values for each bin. We normalized iHS by DAF bin (see equation above). For the sake of comparison, we used exactly the same procedure as for DIND. We grouped mutations by DAF bin (from 0 to 1, in increments of 0.025) and extracted the top 1% of absolute iHS values for each bin. In accordance with a previous study (Voight et al. 2006), we evaluated power on the basis of the proportion of extreme iHS or DIND values in each window. We determined the critical values defining 1% of the 100-kb regions with the highest proportion of extreme iHS or DIND values in 10^4 neutral simulations (equivalent to an FPR of 1%) for each simulated population. The power of each test to detect selection (i.e., either background selection or various regimes of positive selection) was then calculated as the proportion of simulations under selection effectively detected by this procedure (i.e., the percentage of simulations presenting proportions of extreme iHS or DIND values above the threshold defined for an FDR of 1%).

Genome-Wide Calculation of iHS and DIND and Identification of Outliers

To calculate iHS and DIND for each SNP of the WGS data sets analyzed, we first determined the ancestral and derived state of each mutation (see above). However, as these statistics are extremely sensitive to the misspecification of derived states, we calculated iHS and DIND only when the derived state was determined unambiguously. If the regions in which iHS and DIND were calculated overlapped with long gaps (>200 kb), the resulting statistics were excluded from the analysis. We carried out these calculations for 86.67%, 87.8%, and 90.71% of the mutations of the 1000G Pilot, 1000G Phase 1, and the CG data sets, respectively. Because the power of the iHS and DIND was estimated from sets of simulations over 100-kb regions, we calculated iHS and DIND over the same genomic regions of 100 kb surrounding each mutation. This ensures that we obtain values of the two statistics on strictly equivalent regions, in terms of the recombination rate, coverage, and AFS of mutations for each core SNP. We also calculated iHS and DIND over genomic regions of 1 Mb surrounding each mutation, to assess the sensitivity of our results to window size. Indeed, DIND uses information from the haplotype diversity over the entire window considered, while iHS may use information from only a part of the region concerned (i.e., iHH is computed only when EHH > 0.05, over a region whose length is mainly dependent on the intensity of selection). As previously described (Voight et al. 2006), we then extracted the 1% most extreme iHS and DIND values by using bins of DAFs (from 0 to 1, in increments of 0.025) and considered these extreme values (outliers) as potential targets of positive selection. For the identification of regions under positive selection, we focused on the degree of clustering of outliers (Voight et al. 2006). We quantified signal strength by determining the proportion of outliers recorded per 100-kb window. We binned the windows by SNP density and considered the 1% of windows with the highest proportion of outliers in each bin to be potentially under positive selection.

Enrichment in SNP Functional Classes and Resampling Method

We calculated the enrichment of genic and nonsynonymous SNPs among iHS and DIND outliers, by logistic regression, controlling for the genomic variation of certain confounding factors (Kudaravalli et al. 2009). These potential confounding factors include the coverage observed in the region surrounding an SNP (e.g., the power to detect positive selection is lower in regions with low coverage, see Results), recombination rate (Voight et al. 2006), and SNP density. We therefore retrieved these items of information for each window. The recombination rate was determined from HapMap recombination maps build 36 for the 1000G Pilot data set and HapMap recombination maps build 37 for the CG and 1000G Phase 1 data sets. We calculated the enrichment in genic and nonsynonymous SNPs from the logistic model as follows:

$$\begin{aligned} Logit[I(genic = 1)] = {} & \beta_1 I(TEST_o = 1) \\ & + [\beta_2 Cov + \beta_3 Rec + \beta_4 NbSNP \\ & + \beta_5 Cov * Rec + \beta_6 Rec * NbSNP \\ & + \beta_7 NbSNP * Cov] + \varepsilon, \end{aligned}$$

with $I(\text{genic} = 1)$ being an indicator function equal to 1 if the SNP is located in a genic (nonsynonymous) region and equal to 0 otherwise, $I(\text{TEST}_o = 1)$ being an indicator function equal to 1 if the SNP shows a signal of selection (i.e., is an outlier) and equal to 0 otherwise, Rec being the mean recombination rate calculated in cM/bp, Cov being the mean coverage, and nbSNP being the number of SNPs in the window. The OR, which measures the relative enrichment of genic (nonsynonymous) SNPs among SNPs with selection signals (outliers), was estimated by $\exp(\beta_1)$, defined as follows:

$$OR = \left[\frac{P(genic \mid SEL)}{P(nongenic \mid SEL)}\right]\left[\frac{P(nongenic \mid not\ SEL)}{P(genic \mid not\ SEL)}\right]$$

with SEL being "with selection signal," that is with a significant result in tests for selection ($\text{TEST}_o = 1$). The OR estimated from a logistic regression model incorporating all confounding factors and the interaction terms (see equation above) is denoted by OR_C. The odds ratio is denoted OR for logistic regression models not taking the confounding factors into account.

The *P* values associated with enrichment were obtained from 10,000 independent resamplings, taking into account the LD between SNPs, a source of noise that can increase the frequency of outliers in a given window. For each resampling, we drew nonoverlapping regions of 500 consecutive SNPs and arbitrarily assigned them to the genic class until we reached the number of genic SNPs observed in each

population. We considered the remaining SNPs to be nongenic and calculated the OR for each resampling. To resample nonsynonymous SNPs, we first determined the distribution of the number of nonsynonymous SNPs per windows of 500 SNPs. We next drew nonoverlapping regions of 500 consecutive SNPs, and randomly assigned a number of SNPs to the nonsynonymous class so as to fit the real distribution, until we reached the number of nonsynonymous SNPs in each population. Considering the remaining SNPs to be nongenic, we calculated the OR for each resampling. For the calculation of the *P* values for OR_C, we first applied a linear regression to the iHS and DIND values, taking into account the same confounding factors.

$$STAT = C + \alpha_1 Cov + \alpha_2 Rec + \alpha_3 NbSNP + \alpha_4 Cov * Rec + \alpha_5 Rec * NbSNP + \alpha_6 NbSNP * Cov + \varepsilon$$

We then used the residual values (ε) to extract the outliers, before applying the resampling method.

GeneTrail and GWAS Analysis

We used the GeneTrail online tool (http://genetrail.bioinf.uni-sb.de, last accessed April 9, 2014) to analyze the enrichment of some GO biological functions (Ashburner et al. 2000) among DIND outliers. This made it possible to analyze the overrepresentation of each GO category among the outliers by comparing our sets of genes under positive selection with the human reference gene set. An FDR adjustment was applied to correct for multiple testing, and the significance threshold was fixed at 0.05.

The National Human Genome Research Institute (NHGRI) database (http://www.genome.gov/gwastudies/, last accessed April 9, 2014) summarizes results from all published genome-wide association (GWA) analyses for which the *P* values are below 1.0×10^{-5} (Hindorff et al. 2009). We first filtered the database to remove associated SNPs for which *P* values were greater than 1.0×10^{-7}, retaining a single entry for each SNP-trait association. We then calculated the proportion of GWAS-SNPs among DIND outliers by accounting for LD. To this end, all the SNPs associated with the same trait or disease and with the same outlier/nonoutlier status in a genic region were counted as one association. These genic regions were determined with the "mapped gene" field of the database. We then compared these results with those expected under neutrality (0.01 vs. 0.99). ORs were calculated for all associated SNPs together and by trait and disease category.

Supplementary Material

Supplementary text, supplementary figures S1–S11, and tables S1–S20 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Pierre Boutillier for assistance with data analyses, and Katie Siddle and Eddie Loh for discussions. This work was supported by the Institut Pasteur, the Centre Nationale de la Recherche Scientifique (CNRS), the Ecole Normale Supérieure de Lyon, and by the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement No. 281297.

References

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–1073.

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65.

Akey JM. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.* 19:711–722.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12:1805–1814.

Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52–58.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 25:25–29.

Barreiro LB, Ben-Ali M, Quach H, Laval G, Patin E, Pickrell JK, Bouchier C, Tichit M, Neyrolles O, Gicquel B, et al. 2009. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet.* 5:e1000562.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. *Nat Genet.* 40:340–345.

Barreiro LB, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat Rev Genet.* 11:17–30.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet.* 74:1111–1120.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* 4:e1000083.

Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA. 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* 15: 1553–1565.

Casto AM, Li JZ, Absher D, Myers R, Ramachandran S, Feldman MW. 2010. Characterization of X-linked SNP genotypic variation in globally distributed human populations. *Genome Biol.* 11:R10.

Charlesworth B. 2012. The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the *Drosophila* X chromosome. *Genetics* 191: 233–246.

Charlesworth B, Nordborg M, Charlesworth D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet Res.* 70:155–174.

Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. *Genome Res.* 20:393–402.

Chevin LM, Hospital F. 2008. Selective sweep at a quantitative trait locus in the presence of background genetic variation. *Genetics* 180: 1645–1660.

Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza LL, Feldman MW, Pritchard JK. 2009. The role of geography in human adaptation. *PLoS Genet.* 5:e1000500.

Crawford JE, Lazzaro BP. 2012. Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. *Front Genet.* 3:66.
Crisci JL, Poh YP, Mahajan S, Jensen JD. 2013. The impact of equilibrium assumptions on tests of selection. *Front Genet.* 4:235.
Delaneau O, Marchini J, Zagury JF. 2012. A linear complexity phasing method for thousands of genomes. *Nat Methods.* 9:179–181.
Diamond J, Bellwood P. 2003. Farmers and their languages: the first expansions. *Science* 300:597–603.
Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327:78–81.
Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I. 2002. Identification of a variant associated with adult-type hypolactasia. *Nat Genet.* 30:233–237.
Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851–861.
Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T, et al. 2008. A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum Mol Genet.* 17: 835–843.
Gluckman PD, Hanson MA. 2006. Evolution, development and timing of puberty. *Trends Endocrinol Metab.* 17:7–12.
Gold B, Kirchhoff T, Stefanov S, Lautenberger J, Viale A, Garber J, Friedman E, Narod S, Olshen AB, Gregersen P, et al. 2008. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. *Proc Natl Acad Sci U S A.* 105: 4340–4345.
Granka JM, Henn BM, Gignoux CR, Kidd JM, Bustamante CD, Feldman MW. 2012. Limited evidence for classic selective sweeps in African populations. *Genetics* 192:1049–1064.
Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD. 2011. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A.* 108:11983–11988.
Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH, et al. 2013. Identifying recent adaptations in large-scale genomic data. *Cell* 152:703–713.
Hernandez RD. 2008. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics* 24: 2786–2787.
Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. *Science* 331:920–924.
Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A.* 106:9362–9367.
Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* 307:1072–1079.
Hoggart CJ, Chadeau-Hyam M, Clark TG, Lampariello R, Whittaker JC, De Iorio M, Balding DJ. 2007. Sequence-level population simulations over large genomic regions. *Genetics* 177:1725–1731.
Jin W, Xu S, Wang H, Yu Y, Shen Y, Wu B, Jin L. 2012. Genome-wide detection of natural selection in African Americans pre- and post-admixture. *Genome Res.* 22:519–527.
Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y, Li S, Tang K, Chen H, et al. 2013. Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* 152:691–702.
Kang J, Huang KC, Xu Z, Wang Y, Abecasis GR, Li Y. 2013. AbCD: arbitrary coverage design for sequencing-based genetic studies. *Bioinformatics* 29:799–801.
Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM. 2006. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res.* 16:980–989.
Kelley JL, Swanson WJ. 2008. Positive selection in the human genome: from genome scans to biological significance. *Annu Rev Genomics Hum Genet.* 9:143–160.
Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET, Pritchard JK. 2009. Gene expression levels are a target of recent natural selection in the human genome. *Mol Biol Evol.* 26:649–658.
Laval G, Patin E, Barreiro LB, Quintana-Murci L. 2010. Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS One* 5: e10284.
Li H. 2011. A new test for detecting recent positive selection that is free from the confounding impacts of demography. *Mol Biol Evol.* 28: 365–375.
Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. 2011. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* 21:940–951.
Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol.* 34:816–834.
Maranville JC, Baxter SS, Torres JM, Di Rienzo A. 2011. Inter-ethnic differences in lymphocyte sensitivity to glucocorticoids reflect variation in transcriptional response. *Pharmacogenomics J.* 13: 121–129.
Meimaridou E, Kowalczyk J, Guasti L, Hughes CR, Wagner F, Frommolt P, Nurnberg P, Mann NP, Banerjee R, Saka HN, et al. 2012. Mutations in *NNT* encoding nicotinamide nucleotide transhydrogenase cause familial glucocorticoid deficiency. *Nat Genet.* 44: 740–742.
Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. *Genome Res.* 15:1566–1575.
Oleksyk TK, Smith MW, O'Brien SJ. 2010. Genome-wide scans for footprints of natural selection. *Philos Trans R Soc Lond B Biol Sci.* 365: 185–205.
Osier MV, Pakstis AJ, Soodyall H, Comas D, Goldman D, Odunsi A, Okonofua F, Parnas J, Schulz LO, Bertranpetit J, et al. 2002. A global perspective on genetic variation at the *ADH* genes reveals unusual patterns of linkage disequilibrium and diversity. *Am J Hum Genet.* 71:84–99.
Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19:826–837.
Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464:768–772.
Pritchard JK, Di Rienzo A. 2010. Adaptation—not by sweeps alone. *Nat Rev Genet.* 11:665–667.
Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol.* 20:R208–R215.
Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. *Evolution* 59:2312–2323.
Quintana-Murci L, Clark AG. 2013. Population genetic tools for dissecting innate immunity in humans. *Nat Rev Immunol.* 13: 280–293.
Raj T, Kuchroo M, Replogle JM, Raychaudhuri S, Stranger BE, De Jager PL. 2013. Common risk alleles for inflammatory diseases are targets of recent positive selection. *Am J Hum Genet.* 92:517–529.
Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.
Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. 2007. Genome-wide

detection and characterization of positive selection in human populations. *Nature* 449:913–918.
Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nat Biotechnol.* 26:1135–1145.
Tang K, Thornton KR, Stoneking M. 2007. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.* 5:e171.
Teshima KM, Coop G, Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? *Genome Res.* 16:702–712.
Turchin MC, Chiang CW, Palmer CD, Sankararaman S, Reich D, Hirschhorn JN. 2012. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat Genet.* 44: 1015–1019.
Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci U S A.* 102:18508–18513.
Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72.
Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG. 2005. Measures of human population structure show heterogeneity among genomic regions. *Genome Res.* 15:1468–1476.
Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. *PLoS Genet.* 3:e90.

6.3 Conclusions et discussion

6.3.1 Résumé des résultats et nouveautés

Dans cette étude, nous avons montré que *DIND* et *iHS* sont des statistiques plus puissantes pour détecter les balayages sélectifs récents que les statistiques basées sur le spectre de fréquences alléliques, y compris dans le contexte d'histoires démographiques réalistes correspondant aux diversités génétiques observées dans les populations africaines, asiatiques ou européennes. Bien que de nombreuses études aient exploré les balayages sélectifs à l'échelle du génome, parfois en utilisant des données de séquençage à haut débit, aucune n'avait mesuré précisément les effets des caractéristiques de telles données sur la puissance des statistiques. Notre travail démontre l'insensibilité de *DIND* et *iHS* à une profondeur de séquençage faible, à la sélection d'arrière plan et aux autres modes de sélection positive (adaptation polygénique ou sélection sur variant pré-existant), ce qui en fait de bons outils pour tester l'existence et la fréquence des balayages sélectifs récents à partir de données de séquençage génome entier.

L'application de ces deux statistiques aux données de *The 1 000 Genomes Project* et de *Complete Genomics* a mis en évidence que les régions présentant des valeurs inhabituelles (ou *outliers*) de *DIND*, au contraire de celles riches en outliers d'*iHS*, sont significativement enrichies en SNP géniques et non synonymes. De plus, nos résultats répliquent ceux de travaux précédents portant sur des données de génotypage. Ils recouvrent en effet un certain nombre de régions dans lesquelles des signatures d'évolution adaptative avaient été détectées précédemment comme celle, emblématique, du gène de la lactase *LCT*. Nous avons également montré que les signaux de sélection positive détectés sont enrichis en SNP associés à un certain nombre de phénotypes, en particulier en mutations augmentant le risque de développer des maladies complexes telles que les maladies auto-immunes. Plusieurs facteurs peuvent être invoqués pour expliquer la contre-performance d'*iHS* sur les données de la version pilote *The 1,000 Genome Project* et de *Complete Genomics* : la sensibilité d'iHS aux erreurs d'identification des mutations, sa plus grande sensibilité à la taille de l'échantillon et le fait que la présence de nombreux allèles à faible fréquence est susceptible de réduire la longueur des haplotypes et de diminuer le signal d'*iHS*, alors que *DIND* n'est que marginalement affecté par ceux-ci.

Enfin, si de nombreux travaux ont utilisé la méthode des *outliers* afin de détecter les régions sous sélection positive dans le génome humain (Akey 2009), dans notre étude, nous avons couplé la détection des outliers de *DIND* avec le calcul des enrichissements en outliers dans certaines classes de SNP fonctionnels ce qui nous a permis d'évaluer qu'entre 70 et 100 gènes ont été la cible de balayages sélectifs dans chacune des populations étudiées. Ces résultats, qui corroborent les résultats d'une autre étude utilisant des statistiques différentes (Enard et al. 2014), suggèrent un rôle modéré mais non négligeable des balayages sélectifs sur l'évolution du génome humain.

6.3.2 Intérêts

Notre étude apporte un éclairage nouveau et des précisions sur plusieurs points. D'abord, elle confirme que les mutations associées à l'augmentation du risque de développer certaines maladies, dont notamment des maladies auto-immunes, ont été la cible de la sélection positive au cours de l'évolution humaine (Abadie et al. 2011, Barreiro and Quintana-Murci 2010, Blekhman et al. 2008, Nielsen et al. 2009, Raj et al. 2013, Ramos et al. 2014). L'accumulation de ces résultats est un élément en faveur de l'hypothèse que les maladies auto-immunes des sociétés modernes pourraient en fait être le résultat d'une sélection positive ancienne visant à augmenter les capacités de nos défenses immunitaires à faire face aux infections par des pathogènes. Ces résultats soutiennent la théorie hygiéniste (Rook 2009, Strachan 2000) qui stipule que c'est la diminution de l'exposition des enfants aux pathogènes qui est responsable de l'augmentation des cas d'allergies au cours du XXème et du début du XXIème siècle. Du point de vue de l'évolution, cette théorie implique que l'Homme se serait adapté au cours de l'évolution pour combattre les infections, jusqu'à favoriser des mutations potentiellement délétères en l'absence d'exposition massive aux pathogènes (Sironi and Clerici 2010).

Ensuite, elle montre que les données de séquençage de génomes entiers, même à faible profondeur, sont utilisables pour détecter les événements de balayage sélectif récents dans des populations humaines, tout en soulignant l'importance d'évaluer la puissance et la robustesse des statistiques utilisées dans le contexte de telles données. Nos résultats ouvrent la voie à une étude plus fine de ces événements dans d'autres populations humaines. Ainsi, le séquençage de génomes, même à faible

profondeur, d'un nombre suffisant d'individus suffirait à étudier les événements de balayage sélectif ayant eu lieu dans une population. Avec un maillage géographique suffisant des populations, il serait alors possible de corréler événements de sélection, phénotypes observés et facteurs environnementaux, sur le modèle de ce qui a déjà été fait pour un certain nombre de paramètres climatiques (Hancock et al. 2011). Cela apporterait des informations précieuses pour mieux comprendre comment les populations humaines se sont adaptées aux changements environnementaux auxquels elles ont dû faire face dans leur histoire, et quel a été l'effet de ces modifications sur la diversité génétique et phénotypique humaine.

Enfin, en affirmant que les balayages sélectifs récents ont eu un impact modéré sur le génome humain, notre étude participe au débat sur le rôle de la sélection positive dans l'évolution humaine, et suggère que les données de séquençage couplées à l'utilisation d'outils appropriés pourraient permettre de mieux appréhender la part des variations phénotypiques dues à l'adaptation à des pressions environnementales chez l'homme, mais également chez d'autres espèces.

Chapitre 7

Environnement, génétique et variation des profils de méthylation de l'ADN

7.1 Contexte

Les variations épigénétiques, via leur impact sur la régulation de l'expression des gènes, expliquent une partie de la variabilité phénotypique humaine. Parmi les différentes composantes du paysage épigénétique, la méthylation de l'ADN peut être utilisée comme marqueur pour étudier la régulation de l'expression des gènes, grâce à son potentiel informatif sur l'activité des facteurs de transcriptions (Burger et al. 2013), mais aussi à son accessibilité. Différents travaux ont montré que des facteurs environnementaux peuvent entraîner des modifications des profils de méthylation, mais également que des facteurs génétiques expliquent une part non négligeable de la variation de ces profils. Cependant, l'importance relative des mutations génétiques et de l'environnement sur la diversité épigénétique humaine demeure inexplorée.

Pour aborder cette question, nous avons choisi d'étudier le profil de méthylation de l'ADN de populations de chasseurs-cueilleurs Pygmées (RHG) et d'agriculteurs (AGR) d'Afrique Centrale. La diversité génétique et l'histoire démographique de ces populations ont été largement étudiées. Les AGR et les RHG se sont séparés il y a environ 60 000 ans, puis les RHG se sont séparés en deux groupes, à l'Ouest et à l'Est de la forêt équatoriale il y a 20 000 ans (Batini et al. 2011, Patin et al. 2009, Quintana-Murci et al. 2008, Verdu et al. 2013). Les AGR ont ensuite connu deux événements d'expansion de la taille de leur population, un démarré tout de suite après la séparation entre AGR et RHG, et un autre il y a environ 10 000 ans, alors

que la taille de la population RHG est restée constante avant de traverser un goulot d'étranglement récemment (Aimé et al. 2013, Batini et al. 2011, Patin et al. 2009). Les RHG présentent aujourd'hui des signatures d'admixture avec leurs voisins AGR, le flux génétique ayant commencé il y a environ 1 000 ans (Bryc et al. 2010, Patin et al. 2014, Verdu et al. 2009). L'histoire évolutive de ces populations a également été étudiée, montrant que des gènes impliqués dans l'immunité, le métabolisme, l'olfaction, la perception du goût, la reproduction et la guérison des plaies sont sous sélection positive (Amorim et al. 2015, Lachance et al. 2012). Le phénotype Pygmée résulte également de plusieurs événements de sélection positive (Amorim et al. 2015, Jarvis et al. 2012, Lachance et al. 2012, Perry and Dominy 2009, Perry et al. 2014, Verdu et al. 2013).

Ces populations constituent un modèle idéal pour l'étude de l'impact de l'environnement et de la génétique sur le méthylome. En effet, les RHG vivent traditionnellement en petits groupes mobiles dans la forêt équatoriale et ont un mode de subsistance basé sur la chasse et la cueillette alors que les AGR sont des populations sédentaires ayant historiquement occupé des milieux ouverts comme la savane et les prairies, et qui ont adopté l'agriculture comme mode de subsistance depuis environ 5 000 ans (Hewlett 2014, Perry and Dominy 2009). Suite à l'installation de groupe d'AGR dans la forêt tropicale au cours du dernier millénaire (Hewlett 2014, Oslisly et al. 2013), on trouve aujourd'hui des populations présentant des histoires génétiques et démographiques différentes (AGR et RHG) mais partageant le même habitat (forêt équatoriale), et des populations ayant la même histoire génétique mais vivant dans des milieux différents (AGR urbains et forestiers).

Notre objectif a donc été dans un premier temps de caractériser les variations de profils de méthylation de l'ADN entre ces différentes populations en utilisant la puce Illumina 450, qui donne accès à l'état de méthylation de plus de 480 000 sites sur l'ensemble du génome. Nous avons cherché à déterminer l'effet respectif des différences d'habitat et des différences génétiques et environnementales passées sur les variations de profils de méthylation de l'ADN et à identifier les fonctions biologiques touchées. Nous avons ensuite mesuré l'effet de facteurs génétiques sur la variabilité du méthylome et testé si la sélection positive a ciblé ces mutations. Le but final était de comprendre dans quelle mesure l'environnement et des facteurs génétiques ont un impact sur les variations du méthylome de ces populations.

7.2 Article 2 : The Epigenomic Landscape of African Rainforest Hunter-Gatherers and Farmers

The Epigenomic Landscape of African Rainforest Hunter-Gatherers and Farmers

Maud Fagny[1,2,3], Etienne Patin[1,2], Julia L. MacIsaac[4], Maxime Rotival[1,2], Timothée Flutre[5], Meaghan J. Jones[4], Katherine J. Siddle[1,2], Hélène Quach[1,2], Christine Harmant[1,2], Lisa M. McEwen[4], Alain Froment[6], Evelyne Heyer[7], Antoine Gessain[8], Edouard Betsem[8,9], Patrick Mouguiama-Daouda[10], Jean-Marie Hombert[11], George H. Perry[12], Luis B. Barreiro[13,*], Michael S. Kobor[4,*]& Lluis Quintana-Murci[1,2]

[1]Institut Pasteur, Unit of Human Evolutionary Genetics, 75015 Paris, France; [2]Centre National de la Recherche Scientifique, URA3012, 75015 Paris, France; [3]Université Pierre et Marie Curie, Cellule Pasteur UPMC, 75015 Paris, France; [4]Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute and Department of Medical Genetics, University of British Columbia, BC V5Z 4H4 Vancouver, Canada; [5]INRA, UMR AGAP, 34060 Montpellier, France; [6]IRD-MNHN, Sorbonne Universités, UMR208, 75005 Paris, France; [7]CNRS, MNHN, Université Paris Diderot, Sorbonne Paris Cité, Sorbonne Université, UMR7206, 75005 Paris, France; [8]Institut Pasteur, Unité d'Epidémiologie et Physiopathologie des Virus Oncogènes, 75015 Paris, France; [9]Faculty of Medicine and Biomedical Sciences, University of Yaoundé I, Yaoundé, Cameroon; [10]Laboratoire Langue, Culture et Cognition (LCC), Université Omar Bongo, Libreville, Gabon; [11]CNRS UMR 5596, Université Lumière-Lyon 2, 69007 Lyon, France; [12]Departments of Anthropology and Biology, Pennsylvania State University, University Park, PA 16802, USA; [13]Université de Montréal, Centre de Recherche CHU Sainte-Justine, H3T 1C5 Montréal, Canada

[*]These authors contributed equally to this work.

Correspondence should be addressed to L.Q.M. (quintana@pasteur.fr)

Abstract

The genetic history of African populations is increasingly well documented, yet their patterns of epigenomic variation remain uncharacterized. Moreover, the relative impacts of DNA sequence variation and temporal changes in lifestyle and ecological habitat on the human epigenome remain unknown. Here we generated genome-wide genotype and DNA methylation profiles for 362 African rainforest hunter-gatherers and sedentary farmers. We found that the current habitat and the historical lifestyle of a population have similarly critical impacts on the methylome, but the biological functions affected strongly differ. Specifically, methylation variation associated with recent changes in habitat mostly concerns immune functions, whereas that associated with historical lifestyle primarily affects developmental processes. Furthermore, methylation variation — particularly that correlated with historical lifestyle — shows strong associations with nearby genetic variants that, moreover, are enriched in signals of natural selection. Our work provides new insight into the genetic and environmental factors affecting the epigenomic landscape of human populations over different time scales.

Africa is the birthplace of modern humans and a region of extensive genetic, cultural, environmental and phenotypic diversity[1,2]. Over the past years, the increasing amounts of genomic data available have provided significant insight into African evolutionary history, including the origins of hunter-gatherers, ancient population structure, and patterns of migration and admixture[3-10]. Moreover, these studies have reported evidence of selection targeting gene functions related to changes in environment, diet, and exposure to infectious disease[1]. While the study of epigenetic variation can inform the interplay between the environment and the genome, the epigenomic landscape of African populations remains unexplored.

DNA methylation — an important epigenetic mark that serves as biomarker for variation in gene regulation[11,12] — can be affected by both inherited DNA sequence variation and environmental factors, such as nutrition, exposure to toxic pollutants and social environment[13-16]. Accumulating evidence indicates that a substantial portion of DNA methylation variation is accounted for by genetic variation (methylation QTLs)[14,17-21], which could affect methylation levels through impaired transcription factor binding[11,12]. Although the role of DNA methylation in gene regulation (active or passive) and the mechanisms involved remain controversial, DNA methylation data provide a rich source of information about ongoing gene activity, and thus it can provide insight into gene functions that contribute to phenotypic variation[11,12]. Recent studies have shown that DNA methylation differences exist between major ethnic groups[18,22-24], highlighting the potential contribution of epigenetic modifications to human phenotypic variation. However, these studies have mostly compared urban populations of different continental ancestries, so the relative impacts of DNA sequence variation and temporal changes in lifestyle and habitat on the human methylome remain unknown.

The central African belt provides an ideal setting in which to address this issue, as it hosts the world's largest group of active hunter-gatherers — the rainforest hunter-gatherers (RHGs, traditionally known as "pygmies") — as well as populations that have adopted an agrarian lifestyle (AGRs) over the last 5,000 years[25,26]. In addition to differing in their subsistence strategies, these two groups differ in other *historical* and *recent* aspects of their evolutionary history[1,25]. The *historical* factors relate to differences in demography and habitat. The ancestors of the RHGs and AGRs diverged ~60,000 years ago[27-30] and subsequently experienced population contractions and expansions, respectively[7]. These groups have also historically occupied separate ecological habitats — the ancestors of RHGs the equatorial rainforest while those of AGRs open spaces, such as savannah and grasslands[25,31]. More *recent* changes in the lifestyles and habitats of these groups are also apparent. Many RHG groups still live in the rainforest as mobile bands, whereas AGR populations now occupy primarily rural or urban deforested areas, though some AGR groups have settled in the rainforest over the last millennia[25,31].

In this study, we defined the genome-wide DNA methylation profiles of various populations of RHG and AGR inhabiting the central African belt to first assess the degree of inter-population variation in DNA methylation. We then explored the genomic and functional features of differentially methylated genes to obtain insight into the putative phenotypes involved. Finally, we assessed the contribution of genetic variation to the DNA methylation levels observed, and searched for signals of positive selection targeting genetic variants associated with methylation variation. The integration of these results allowed us to develop a comprehensive framework of how temporal differences in lifestyle and habitat, together with genetic variation, have impacted the epigenomic landscape of human populations.

Results

Population samples and DNA methylation data set. We investigated genome-wide genotype and DNA methylation data from a total of 362 individuals, including a group of RHGs (w-RHG, *n*=112), AGR groups occupying nearby urban deforested habitats (w-AGR, *n*=94), and an AGR group living in a forested region (f-AGR, *n*=61) of the Gabon/Cameroon area (Fig. 1a, Table 1). To compare our results with an independent set of samples, we also studied RHGs and AGRs living in the eastern part of the Central African belt (e-RHG, *n*=47 and e-AGR, *n*=48, from Uganda). The genetic structure of these populations, which we investigated using genome-wide SNP data, reflects their history of population divergence[27-30], with the largest differences being those between RHGs and AGRs, followed by the split between the western and eastern RHG groups (Fig. 1b).

We assessed DNA methylation differences in whole blood-derived samples using the Illumina 450K array, which interrogates more than 485,000 sites across the genome. After normalization and filtering, we retained 365,886 probes, which were evaluated in 352 individuals (Methods). We validated the DNA methylation array findings by bisulfite pyrosequencing of four regions in all samples. We observed a good correlation between methylation levels measured by pyrosequencing and the array (Pearson $R = 0.89$ for cg23053977, $R = 0.88$ for cg08684511, $R = 0.74$ for cg09879458, and $R = 0.90$ for cg18757155, Supplementary Fig. 1a-d). Moreover, the two methods showed good concordance (Supplementary Fig. 1e-h), with pyrosequencing data presenting a larger range of values at both completely methylated sites (cg23053977) and completely unmethylated sites (cg18757155) (Supplementary Fig. 1e,h). Samples showed expected DNA methylation profiles across genomic regions, with sites near gene promoters being largely unmethylated (Supplementary Note 1, Supplementary Fig. 2).

Differences in genome-wide DNA methylation profiles. We initially compared DNA methylation variation between populations differing in genetic background, historical lifestyle and current habitat — the RHG and AGR groups living in the rainforest and rural/urban areas, respectively (Fig. 1c). We adjusted DNA methylation values for age and sex, and accounted for heterogeneity in blood cell composition (Methods; Supplementary Notes 2 and 3, Supplementary Figs 3 and 4). Principal component analysis (PCA) of DNA methylation data clearly separates the RHG and AGR groups on PC1, in both western ($P = 3.3\times10^{-9}$) and eastern ($P = 2.7\times10^{-9}$) central Africa (Fig. 2a,b and Supplementary Fig. 5). We identified 25,913 differentially methylated sites (DMS) (8,823 genes) between w-RHG and w-AGR, and 19,429 DMS (6,294 genes) between e-RHG and e-AGR (FDR < 0.01). Comparing the western and eastern settings, we detected a strong overlap, with 6,852 sites (2,531 genes) differentially methylated in the same direction — corresponding to 96% of the overlapping DMS (resampling $P < 10^{-7}$). These findings attest to strong, shared differences in DNA methylation between RHG and AGR groups, regardless of their geographic location.

Impact of temporal changes in habitat and lifestyle on DNA methylation. To distinguish the respective effects on DNA methylation of recent changes in habitat from historical differences in lifestyle and genetics of these groups, we next compared populations with a common historical lifestyle and genetic background but different recent habitats: the forest f-AGR and the urban w-AGR (Fig. 1c). The observed patterns of DNA methylation variation were accounted for primarily by the habitat in which the populations live (PC1 $P=2.7\times10^{-3}$; Fig. 2c), highlighting the important role of current habitat in determining global DNA methylation profiles. We found 5,765 DMS (3,570 genes) between the two groups, which we termed "*recent* DMS". The differential methylation in the same direction of 3,338 of these

recent DMS (corresponding to 99% of the overlapping DMS, resampling $P < 10^{-7}$; 2,166 genes) between the more distantly related w-RHG and w-AGR provides strong evidence in favour of the methylation status at these shared DMS being determined by recent changes in habitat independently of genotypic differences.

Focusing on populations with different historical lifestyles and genetic backgrounds but with the same current habitat (f-AGR and w-RHG in the central African rainforest, Fig. 1c), PCA also tended to separate the samples with respect to their population identity (PC1 $P=3\times10^{-4}$; Fig. 2d). We found 4,054 DMS (2,130 genes) between these groups, which we termed "*historical* DMS". These *historical* DMS showed no significant overlap with the *recent* DMS described above (only 52 DMS were shared). Notably, the proportion of DMS for which mean DNA methylation levels differed strongly between populations (i.e., $\Delta\beta$ values > 5%) was higher for *historical* than for *recent* DMS ($P < 10^{-16}$, Supplementary Fig. 6). The set of *historical* DMS identified reflect DNA methylation variation related to the historical differences in lifestyle and habitat characterizing the RHG and AGR groups.

Genomic features of differentially methylated regions. To understand the putative functional implications of DMS, we first localized them across distinct genomic regions. We found that *recent* DMS were enriched in sites located in gene bodies and distal promoters, while *historical* DMS were preferentially located around the TSS, 5'-UTR and 1st exon regions (Supplementary Fig. 7a,c). We next mapped DMS to histone modification peaks from PBMCs described by the ENCODE project[32]. We found that both *recent* and *historical* DMS mapped in excess to H3K4me1 modification peaks (32% for both DMS sets *vs*. 20% expected) (Supplementary Fig. 7b,d). Notably, the *recent* DMS that were upmethylated in f-

AGR were further enriched in H3K4me3 peaks (57% *vs*. 27%), while the *historical* DMS that were upmethylated in w-RHG were enriched in H3K27me3 (32% *vs*. 12%).

Finally, we explored the co-localization of DMS with transcription factor (TF) binding sites (Methods; Supplementary Table 1). We found that *recent* DMS were significantly enriched in binding sites of TF related to cell differentiation, proliferation and development, but also to immunity regulation (NFIL3, IRF1 and SPIB) and fatty acid storage and glucose metabolism (HNF1A, PPARG and NR1H2::RXRA). Conversely, *historical* DMS, particularly those that were upmethylated in RHG, were preferentially overlapping binding sites of TF involved in developmental processes (TFAP2A and GATA2). Collectively, these findings indicate that *recent* and *historical* DMS not only correspond to independent sets, but also they are located in distinct genomic regions that contain different TF binding sites, suggesting that they are associated to regulatory features related to different biological functions.

Biological functions of recent and historical DMS strongly differ. We investigated the relevance of *recent* and *historical* DMS for explaining phenotypic diversity, by exploring the gene ontology categories associated with differentially methylated genes in each set. We found striking differences in the biological functions involved. Genes containing *recent* DMS were enriched in categories almost exclusively related to immunity, such as immune response, viral process and cell surface receptor signalling pathways (Fig. 2e, Supplementary Table 2).

Conversely, genes overlapping *historical* DMS were enriched in functions largely associated with development, including multicellular organismal development, anatomical structure development, or growth factor binding (Fig. 2f, Supplementary Table 3). We found that 1,302 *historical* DMS (699 genes) overlapped with the DMS detected in western (w-AGR vs. w-

RHG) and eastern (e-RHG vs. e-AGR) comparisons, in the same direction (corresponding to 99% of the overlapping DMS, resampling $P < 10^{-7}$), despite the splitting of the RHG groups ~20,000 years ago[28,29]. This common set of *historical* DMS was again enriched in functions primarily related to development (Supplementary Table 4). We thus identified a gene set in which epigenomic variation reflects differences in the lifestyle and habitat, as well as in genetic background, of RHGs and AGRs, regardless of their geographic location.

Genetic contribution to DNA methylation variation. To assess the contribution of genetic variation to the DNA methylation levels, we mapped methylation QTLs (meQTLs), focusing our analyses on SNPs located in *cis* within a 200 kb window around the target site (Methods; Supplementary Fig. 8). We identified 46,017 DNA methylation sites (~13% of all sites) associated with a nearby meQTL, in at least one population, with a FDR set to 1%. The majority of meQTLs (~90%) were shared across populations, with only 1,289 and 502 meQTLs detected exclusively in the RHG and AGR groups, respectively. Such extensive sharing of meQTLs reflects the closer genetic proximity of the populations studied here and the use of a different cellular model, with respect to previous studies[22,24] (Supplementary Table 5, Supplementary Fig. 9 and Supplementary Note 4).

We then tested the potential enrichment of differentially methylated regions in associations with genotype variants, with respect to all DNA methylation sites. We found a moderate enrichment in DMS characterizing the western (17%, OR=1.6, SE=0.02; resampling $P < 10^{-7}$) and eastern comparisons (14%, OR=1.2, SE=0.02; resampling $P = 2.5\times10^{-2}$), where populations differ in both historical and recent lifestyles and habitats (Fig. 3a). Furthermore, *historical* DMS were strongly enriched in meQTLs (33%, OR=3.6, SE=0.03; resampling $P < 10^{-7}$), whereas *recent* DMS were depleted in these associations (10%, OR=0.79, SE=0.04;

resampling $P < 10^{-7}$). Notably, the proportion of the variance of DNA methylation accounted for by meQTLs (R^2) was higher for meQTLs associated with *historical* DMS (~11%) than for meQTLs related to *recent* DMS (6.6%), the R^2 values obtained being significantly higher and lower, respectively, than for all meQTLs (Fig. 3b). Consistent with all our previous observations, *historical* DMS are more strongly associated with genotypic differences, which have also a larger effect, than the remaining sets of DMS.

Two scenarios can explain the observed associations between *historical* DMS and DNA sequence variants. In most cases, DNA methylation differences were accounted for by meQTLs detected in all populations but with differences in allelic frequency between the RHG and AGR groups (Fig. 3c-e, Supplementary Fig. 10a-c). More rarely, genetic variants appeared to correlate with DNA methylation differences only in some populations, indicating interactions with other genetic variants and/or the environment (G×G or G×E interactions) (Fig. 3f, Supplementary Fig. 10d).

Signals of positive selection targeting meQTLs. Finally, we explored the adaptive significance of meQTLs using two statistics — F_{ST} and iHS — that detect positive selection signals based on population differentiation[33] and haplotype homozygosity[34], respectively (Methods). We found that meQTLs were significantly enriched in high F_{ST} values with respect to the remainder of the genome (Cochran–Mantel–Haenszel $P < 1.2\times10^{-5}$, stratified by derived allele frequencies), in population comparisons involving the RHG and AGR groups (Fig. 4a). Likewise, the distributions of $|iHS|$ values were significantly skewed towards higher values for meQTLs among AGR groups (Mann-Whitney U $P < 5.3\times10^{-8}$), suggesting more recent events of positive selection (Fig. 4b). Collectively, these findings suggest that

positive selection has targeted DNA sequence variants that influence — directly or indirectly — variation in DNA methylation.

Discussion

The dissection of the epigenetic landscape of African rainforest hunter-gatherers and sedentary farmers shows that a population's current habitat and historical lifestyle have both critical impacts on the patterns of methylation variation. However, we demonstrate that the biological functions affected differ strongly depending on the timing of these events. Recent changes in habitat, such as those experienced by agriculturalist populations living in urban/rural areas or in the rainforest, affect principally immune functions. This suggests functional links between DNA methylation variation and host responses to recent changes in habitat exposure, as observed for gene expression in Moroccan populations[35].

In contrast, when comparing rainforest hunter-gatherers and farmers who share the same forest environment — a setting that minimizes the effects that recent environmental changes have had on methylation — we find that methylation differences related to historical factors mostly concern genes with functions in developmental processes. Furthermore, differentially methylated regions that correlate with historical factors are strongly associated with genetic variants, the frequency of which differs between hunter-gatherer and farmer groups. This is the case, for example, for genes such as *IGF2BP2*, *HOXC6* and *ZNF492* (Fig. 3c-e), which have been associated with height[36] — the most iconic adaptive phenotype characterizing RHGs[4,37,38] — age at menarche[39], type-2 diabetes[40], bone mineral density[41], and gene-diet interactions[42]. We also observe cases of population-specific effects of DNA methylation variation, such as for *PRKCZ* — also associated with height[36] — that was hypomethylated in rainforest hunter-gatherers and under genetic control only in this group (Fig. 3f).

In summary, this study increases our understanding of the relative impacts that population genetic variation and differences in lifestyles and ecologies have on the human epigenome, and illustrates the utility of DNA methylation as a marker to track variation in regulatory activity following environmental change. Furthermore, our findings suggest that populations can initially respond to environmental challenges via epigenetic changes, uncoupled from variation in the DNA sequence, with the adaptive phenotype increasingly being achieved via genetic changes as time passes. We thus provide a basis for further experimental and theoretical studies assessing the role of epigenetic mechanisms in human adaptation over different time scales.

Methods

Population Samples. We studied peripheral whole blood DNA from a total of 381 samples, corresponding to 362 individuals and 19 replicate samples, from seven populations located across the central African belt (Fig. 1a, Table 1). These populations can be divided into two main groups: rainforest hunter-gatherer (RHG) populations, historically known as "pygmies", who have traditionally relied on the equatorial forest for subsistence and who live close to, or within, the forest; and agricultural (AGR) populations, living either in rural/urban deforested regions or in forested habitats in which they practice slash-and-burn agriculture. The w-RHG sample consisted of 112 Baka from Minvoul (Gabon) and the regions of Oveng-Djoum, Lomié-Messok, and Salapoumbe (Cameroon). Given the highly similar genetic and methylation profiles of the Baka individuals from Cameroon and Gabon (Fig. 1b, Supplementary Fig. 5), and their residence in the same ecological habitat (Table 1), we pooled these samples in a single group. The e-RHG sample consisted of 47 unrelated Batwa from the surroundings of the Bwindi Impenetrable Forest in southwest Uganda, all of whom were born in the forest[38]. The w-AGR sample contained 55 Nzebi from Libreville (Gabon) and 39 Fang from Yaoundé (Cameroon). Again, based on the similarity of their genetic and methylation profiles (Fig. 1b, Supplementary Fig. 5) and habitats (Table 1), these samples were merged into a single group. The e-AGR sample contained 48 Bakiga from the surroundings of the Bwindi Impenetrable Forest in southwest Uganda[38]. We also analysed an AGR sample of 61 Nzime from Messok (Cameroon) (referred to as f-AGR), from the same forest habitat as the w-RHG sample. Further details about the modes of subsistence of these populations, their habitats and sample sizes, before and after filtering, are provided in Table 1. Informed consent was obtained from all participants and from both parents of any participants under the age of 18. Ethical approval for this study was obtained from the institutional review boards of

Institut Pasteur, France (RBM 2008-06 and 2011-54/IRB/3), Makerere University, Uganda (IRB 2009-137) and University of Chicago, USA (16986A).

Genotyping Data. Of the 362 individuals included in this study, 191 had already been genotyped by Illumina Omni1 in two previous studies[7,38]. This group consisted of 46 w-RHG, 15 e-RHG, 29 w-AGR, 31 e-AGR and 21 f-AGR individuals from ref.[7], and 34 e-RHG and 15 e-AGR individuals from ref.[38]. The remaining 171 samples — 105 w-RHG, 26 w-AGR and 40 f-AGR individuals — were genome-wide genotyped using the Illumina OmniExpress for 719,665 SNPs. We filtered out 7,120 SNPs on the basis of their physical location (i.e., those on the Y-chromosome and SNPs unmapped on dbSNP build 37), problematic genotype clusters in GenomeStudio (Illumina, San Diego) based on a GenTrain score < 0.35, and SNP call rate <95%. We also filtered out two w-RHG individuals with a call rate <95% and eight individuals presenting cryptic relatedness (i.e., kinship coefficient > 0.15 with another individual), with the KING program[43]. We phased the 191 Omni1 individuals with SHAPEIT2[44] and imputed missing SNPs in the OmniExpress dataset, using the Omni1 dataset as a reference, with IMPUTE2[45]. Five samples (4 w-RHG and 1 f-AGR) with a call rates <95% after imputation were removed. After filtering out low-quality imputed SNPs and SNPs with call rate <95% after imputation, we obtained a final set of genotypes at 876,886 SNPs for 347 individuals, comprising 98 w-RHG, 94 w-AGR, 60 f-AGR, 47 e-RHG and 48 e-AGR individuals. We then had to remove another two individuals because of their methylation profiles (see the "DNA methylation data processing" section), yielding a final dataset of 345 individuals for whom we had both genotype and methylation data.

Genome-Wide DNA Methylation Analysis. Genome-wide DNA methylation data at more than 485,000 sites was obtained using an Infinium® HumanMethylation450 BeadChip.

Bisulfite conversion of 750 ng of genomic DNA was performed with the EZ DNA Methylation™ Kit. Successful conversion was confirmed by methylation-specific PCR prior to proceeding with subsequent steps of the Infinium assay protocol. The bisulfite-converted genomic DNA was isothermally amplified at 37°C for 22 h, enzymatically fragmented, purified and hybridized with the HumanMethylation450 BeadChip at 48°C for 18 h. Each BeadChip was then washed to remove any un-hybridized or non-specifically hybridized DNA. Labelled single-base extension was performed with bead-bound probes hybridized to the DNA, and the hybridized DNA was removed. The extended probes were stained with multiple layers of fluorescence, and the BeadChip was then coated with a proprietary solution and scanned with the Illumina® iScan system. Raw data were processed with Genome Studio™ Methylation Module software.

Targeted Pyrosequencing. Bisulfite PCR-pyrosequencing assays were designed with PyroMark Assay Design 2.0 (Qiagen). The regions of interest (*RORA* cg09879458, enhancer region cg23053977, *ADAM28* cg18757155 and *COL23A1* cg08684511) were amplified by PCR, using the HotstarTaq DNA polymerase kit (Qiagen) as follows: 15 minutes at 95°C (to activate the *Taq* polymerase), 45 cycles of 95°C for 30 s, 58°C for 30 s, and 72°C for 30 s, with a final five-minute extension step at 72°C. For pyrosequencing, a single-stranded DNA was prepared from the PCR product with the Pyromark™ Vacuum Prep Workstation (Qiagen), and sequencing was performed with sequencing primers on a Pyromark™ Q96 MD pyrosequencer (Qiagen). Methylation levels were calculated for each CpG dinucleotide with Pyro Q-CpG software (Qiagen). The primer sequences are listed in Supplementary Table 6.

DNA Methylation Data Processing. In total, 381 samples were hybridized with the HumanMethylation450 array, including 362 unique samples and 19 technical replicates. We

removed probes that potentially cross-hybridize[46], those on the X and Y chromosomes, and those containing SNPs at a frequency higher than 1%. The list of SNPs was based on (i) our own genotyping dataset for more than 876,886 SNPs genome-wide (see "Genotyping data" section), and (ii) a whole-genome sequencing dataset for 20 w-AGR and 20 w-RHG individuals from this collection, sequenced at an average depth of coverage of 5.6× (17,080,726 SNPs, unpublished data). Following this filtering process, 365,886 of the original 485,512 sites on the array were retained. We calculated methylation levels from raw data, using the R bioconductor lumi package. The M-value has been shown to provide better detection sensitivity than β-values at extreme levels of modification[47]. We therefore used the M-value unless otherwise stated. M-values were then adjusted for background and color bias with lumi, and quantile-normalized. We corrected for technical differences between Type I and Type II assay designs, by performing subset-quantile within array normalization (SWAN) on M-values with the R bioconductor minfi package[48]. PCA showed that a batch effect explained part of the variance (Kruskal-Wallis *P*-value of 8.35×10^{-55} for PC2) of the normalized data, and we used the ComBat function from the sva bioconductor package to correct for this effect[49]. Two samples (1 w-RHG and 1 f-AGR) were removed because they presented a clear excess of hemi-methylated sites.

Accounting for Heterogeneity in Cell Subtypes and Age. We accounted for potential cellular heterogeneity in whole blood samples by estimating cell composition in all individuals. We used a reference-based method in which the DNA methylation signature of each of the principal types of immune cells (granulocytes, monocytes, B cells, $CD4^+$ T cells, $CD8^+$ T cells and NK cells) was used to predict the proportions of these cells types in unfractionated whole blood[50]. Predictions for white blood cell types were obtained by applying the "estimateCellCounts" function of the minfi package[48] to normalized β-values.

This function was modified slightly to accept a matrix of β-values rather than an RGSet object. The resulting estimated cell counts were rescaled to 1. Estimated cell subtype heterogeneity across populations was then used to adjust M-values for subsequent analyses, including principal component analyses, the estimation of differentially methylated sites and the mapping of methylation quantitative trait loci. We also determined the relative proportions of various cell subtypes ($CD4^+$ T cells, $CD8^+$ T cells, B cells and NK cells) among the peripheral blood mononuclear cells (PBMCs) of 35 e-RHG and 31 e-AGR subjects, by fluorescence-activated cell sorting (FACS). Ages were imputed from methylation data for all samples, with an elastic net regression model, as previously described[51], and the imputed ages were compared with the ages declared, when available (Supplementary Note 2). Imputed ages were then used to adjust M-values in all populations, for all analyses.

Determination of Differentially Methylated Sites. Sites differentially methylated between populations (DMS) were identified statistically, by fitting a linear regression model for each site (M-values ~ population + sex + age + cell type proportions + error), and applying empirical Bayes smoothing to the standard errors, with the R bioconductor limma package[52]. Sites with a Benjamini & Hochberg adjusted $P < 0.01$ were considered to be differentially methylated. To define the amplitude of DMS, we used different criteria: a Benjamini & Hochberg adjusted P-value lower than 0.01 and a difference in mean methylation level between the two populations of more than 2%, 5% or 10%. For this analysis, methylation level was determined as the ratio of methylated probe intensity to overall intensity, the β-value[47]. We extracted the overlaps between different DMS sets and calculated the P-values measuring the probability of these overlaps being obtained by chance, using 10^7 resamplings. DNA methylation levels at targeted sites are strongly correlated within regions of about 2,000

bp[18]. Thus, for each DMS list, we randomly resampled the same number of sites from all 365,886 sites, taking into account the distance between the DMS.

Genic Distribution of DMS and Enrichment in Histone Marks and TFBS. We analyzed the enrichment in target sites of particular genomic regions, by calculating an odds ratio, defined as follows:

$$OR = \left[\frac{P(R|DMS)}{P(not\ R|DMS)}\right]\left[\frac{P(not\ R|not\ DMS)}{P(R|not\ DMS)}\right]$$

with R being "in the region".

Genic regions were defined according to the UCSC_REFGENE_GROUP column from the Illumina HumanMethy450 annotation: distal promoter (from 1,500 to 200bp upstream from the TSS), proximal promoters (less than 200bp upstream from the TSS), 5'UTR, 1st Exon, Gene Body and 3'UTR. Histone modification peak data for H3k4me1, H3K4me3, H3K9me3 and H3K27me3 in PBMCs were downloaded from the ENCODE website (http://genome.ucsc.edu/ENCODE/). A site was considered to be colocalised with a histone modification mark if it fell into the region defined as a "narrow peak" (FDR of 0.01). Transcription factor (TF) binding sites affinity scores for sequences of 30bp around each methylation site were obtained using the TRAP software[53] and the position weight matrix of 85 human TFs from JASPAR[54]. For each TF, a site was considered to have a high affinity if it fell into the top 5th percentile of the score distribution. *P*-values for enrichment in genomic positions, histone marks or TFBS among *recent* and *ancient* DMS were obtained using a χ^2-test.

Biological Functions of Differentially Methylated Genes. We extracted all differentially methylated genes, defined as genes carrying at least one DMS. We used the goseq R bioconductor package to perform an analysis of the over-representation of Gene Ontology (GO) categories[55] among differentially methylated genes. We fed the number of probes corresponding to each gene into the Probability Weighting Function of the goseq package. As not all the genes of the genome are represented on the Illumina HumanMethy450 BeadChip, our reference set in the over-representation analysis consisted of the 19,672 genes retained after filtering for which we had data. DMSs were significantly enriched in a given category if the FDR-adjusted *P*-value was lower than 0.05.

Mapping of Methylation Quantitative Trait Loci (meQTLs). We identified meQTLs with a Bayesian statistical framework implemented in the eQtlBma package, which was specifically designed for the detection of QTLs jointly in multiple subgroups[56]. We filtered out SNPs with an allele frequency below 10% in all populations. Age, sex, the proportions of the various cell types and the first PC for genotyping data were used as covariates in the linear model. We then estimated the genome-wide weight of each configuration (Supplementary Table 5) using eqtlbma_hm and the default grids provided by the eQtlBma package as a priori for the hierarchical model. The probability of a methylation site having no meQTL ($\pi 0$) was estimated by the EBF method[57], and various posterior probabilities were calculated with eqtlbma_avg_bfs. We then extracted all the methylation sites with at least one meQTL at an FDR of 0.01, as previously described[58]. We identified the best associated SNPs, defined as all SNPs for which the sum of posterior probabilities for being the best associated SNP, assuming that the site was associated with only one SNP, was at least 0.85. For most sites with several SNPs associated with high posterior probabilities, the best configurations (i.e. the combination of populations in which the SNP was a meQTL) were identical for all the SNPs.

In the 2,027 cases in which there were at least two configurations, the best configuration was chosen by looking directly at the association. The 156 cases for which there were more than two different configurations were discarded from the list of significant meQTL-associated sites.

Detection of Positive Selection. To detect mutations presenting signals of positive selection, we used the population differentiation-based F_{ST}[33], and the haplotype-based, integrated haplotype score (iHS)[34] for our set of 876,886 SNPs. Population pairwise AMOVA-based F_{ST} values were computed for w-RHG *vs*. w-AGR, e-RHG *vs*. e-AGR, f-AGR *vs*. w-AGR, and w-RHG *vs*. f-AGR. To measure the enrichment in high F_{ST} values among meQTLs, we used a Cochran–Mantel–Haenszel test to compare the proportions of meQTLs with high F_{ST} values with those of the non-meQTL SNPs, stratifying data by bin of derived allele frequencies (from 0 to 1, in 0.05 steps). iHS were computed for all SNPs, and normalized by bin of derived allele frequencies (DAF, from 0 to 1, in 0.025 steps) in each of the five populations separately (w-RHG, w-AGR, f-AGR, e-RHG and e-AGR). Ancestral states of the SNPs were determined using the sequence provided by the 1000 Genomes Project[59]. We used a Mann-Whitney U test to compare DAF-matched iHS distributions among meQTL and non-meQTL SNPs. For both F_{ST} and iHS analysis, we filtered out SNPs with LD r^2 values higher than 0.8 in each pair of populations or in each population separately, respectively, using plink[60].

References

1 Campbell, M. C. & Tishkoff, S. A. The evolution of human genetic and phenotypic variation in Africa. *Curr Biol* **20**, R166-173 (2010).

2 Campbell, M. C., Hirbo, J. B., Townsend, J. P. & Tishkoff, S. A. The peopling of the African continent and the diaspora into the new world. *Curr Opin Genet Dev* **29**, 120-132 (2014).

3 Schlebusch, C. M. *et al.* Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* **338**, 374-379 (2012).

4 Lachance, J. *et al.* Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse african hunter-gatherers. *Cell* **150**, 457-469 (2012).

5 Schuster, S. C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943-947 (2010).

6 Pickrell, J. K. *et al.* The genetic prehistory of southern Africa. *Nat Commun* **3**, 1143 (2012).

7 Patin, E. *et al.* The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. *Nat Commun* **5**, 3163 (2014).

8 Henn, B. M. *et al.* Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc Natl Acad Sci U S A* **108**, 5154-5162 (2011).

9 Veeramah, K. R. *et al.* An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. *Mol Biol Evol* **29**, 617-630 (2012).

10 Bryc, K. *et al.* Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci U S A* **107**, 786-791 (2010).

11 Schubeler, D. Function and information content of DNA methylation. *Nature* **517**, 321-326 (2015).

12 Pai, A. A., Pritchard, J. K. & Gilad, Y. The Genetic and Mechanistic Basis for Variation in Gene Regulation. *PLoS Genet* **11**, e1004857 (2015).

13 Kaminsky, Z. A. *et al.* DNA methylation profiles in monozygotic and dizygotic twins. *Nature Genet* **41**, 240-245 (2009).

14 Lam, L. L. *et al.* Factors underlying variable DNA methylation in a human community cohort. *Proc Natl Acad Sci U S A* **109 Suppl 2**, 17253-17260 (2012).

15 Feil, R. & Fraga, M. F. Epigenetics and the environment: emerging patterns and implications. *Nat Rev Genet* **13**, 97-109 (2011).

16 Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477-481 (2013).

17 Gibbs, J. R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet* **6**, e1000952 (2010).

18 Bell, J. T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol* **12**, R10 (2011).

19 Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2**, e00523 (2013).

20 Banovich, N. E. *et al.* Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet* **10**, e1004663 (2014).

21 Zhang, D. *et al.* Genetic control of individual differences in gene-specific methylation in human brain. *Am J Hum Genet* **86**, 411-419 (2010).

22 Fraser, H. B., Lam, L. L., Neumann, S. M. & Kobor, M. S. Population-specificity of human DNA methylation. *Genome Biol* **13**, R8 (2012).

23 Heyn, H. *et al.* DNA methylation contributes to natural human variation. *Genome Res* **23**, 1363-1372 (2013).

24 Moen, E. L. *et al.* Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits. *Genetics* **194**, 987-996 (2013).

25 Hewlett, B. S. *Hunter-Gatherers of the Congo Basin : Culture, History and Biology of African Pygmies*. (Transaction Publishers, 2014).

26 Perry, G. H. & Dominy, N. J. Evolution of the human pygmy phenotype. *Trends Ecol Evol* **24**, 218-225 (2009).

27 Quintana-Murci, L. *et al.* Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proc Natl Acad Sci U S A* **105**, 1596-1601 (2008).

28 Verdu, P. *et al.* Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Curr Biol* **19**, 312-318 (2009).

29 Patin, E. *et al.* Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet* **5**, e1000448 (2009).

30 Batini, C. *et al.* Insights into the demographic history of African Pygmies from complete mitochondrial genomes. *Mol Biol Evol* **28**, 1099-1110 (2011).

31 Oslisly, R. *et al.* Climatic and cultural changes in the west Congo Basin forests over the past 5000 years. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **368**, 20120304 (2013).

32 ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

33 Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat Rev Genet* **10**, 639-650 (2009).

34 Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol* **4**, e72 (2006).

35 Idaghdour, Y., Storey, J. D., Jadallah, S. J. & Gibson, G. A genome-wide gene expression signature of environmental geography in leukocytes of Moroccan Amazighs. *PLoS Genet* **4**, e1000052 (2008).

36 Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-1186 (2014).

37 Jarvis, J. P. *et al.* Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS Genet* **8**, e1002641 (2012).

38 Perry, G. H. *et al.* Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proc Natl Acad Sci U S A* **111**, E3596-3603 (2014).

39 Perry, J. R. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92-97 (2014).

40 Mahajan, A. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* **46**, 234-244 (2014).

41 Estrada, K. *et al.* Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat Genet* **44**, 491-501 (2012).

42 Figueiredo, J. C. *et al.* Genome-wide diet-gene interaction analyses for risk of colorectal cancer. *PLoS Genet* **10**, e1004228 (2014).

43 Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873 (2010).

44 Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-181 (2012).

45 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).

46 Price, M. E. *et al.* Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* **6**, 4 (2013).

47 Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).

48 Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol* **13**, R44 (2012).

49 Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882-883 (2012).

50 Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).

51 Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol* **14**, R115 (2013).

52 Smyth, G. K. in *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (eds R. Gentleman *et al.*) 397-420 (Springer, 2005).

53 Thomas-Chollier, M. *et al.* Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nat Protoc* **6**, 1860-1869 (2011).

54 Bryne, J. C. *et al.* JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res* **36**, D102-106 (2008).

55 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).

56 Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet* **9**, e1003486 (2013).

57 Wen, L. Robust Bayesian FDR Control with Bayes Factors. *arXiv:1311.3981 [stat.ME]* (2013).

58 Newton, M. A., Noueiry, A., Sarkar, D. & Ahlquist, P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155-176 (2004).

59 Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

60 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).

Acknowledgements

We thank Vincent Colot, Etienne Danchin, Jean-Philippe Fortin, Tatiana Giraud, Aurélie Labbe, Guillaume Laval, and Carla Saleh for feedback on data analyses and reading of the manuscript. We are grateful to all the study participants for their generous contributions of DNA. This study was funded by the Institut Pasteur, the CNRS, a CNRS "MIE" (Maladies Infectieuses et Environnement) Grant, and a Foundation Simone & Cino del Duca Research Grant (L.Q.-M.), and the Canadian Institute for Advanced Research (CIFAR) (M.S.K.). M.J.J was supported by a Mininig for Miracles post-doctoral fellowship from the Child and Family Research Institute. L.B.B is supported by the Canada Research Chairs Program. M.S.K. is the Canada Research Chair in Social Epigenetics and a Senior Fellow of CIFAR.

Author contributions

L.Q.-M. conceived and supervised the study. M.F. designed the analysis strategy and analysed the data, with input from E.P., M.R., M.J.J., M.S.K., L.B.B and L.Q.-M. T.F., M.R. M.J.J. and K.J.S. provided support for the analysis strategy and statistical methods. J.L.M., L.M.M. and M.S.K. contributed DNA methylation data and performed targeted pyrosequencing. H.Q. and C.H. assisted with the genetic analyses. A.F., E.H., A.G., E.B., P.M-D., J.-M.H., G.H.P, and L.B.B contributed to sample collection. L.B.B contributed FACS data. M.F and L.Q.-M. wrote the manuscript, with input from all authors.

Additional information

Accession Numbers. The data reported have been deposited in the European Genome-Phenome Archive, www.ebi.ac.uk/ega/home (genotyping data accession no. EGAS00001000605, EGAS00001000908, and EGAS00001001066, and DNA methylation data accession no. EGAS00001001066).

Competing Financial Interests. The authors declare no competing financial interests.

Figure Legends

Figure 1 | Study design and genetic structure of rainforest hunter-gatherers and farmers. (**a**) Geographic location of the sampled rainforest hunter-gatherer (RHG) and farmer (AGR) populations (Table 1). (**b**) Principal component analysis (PCA) of the genotype data for the study populations, based on 456,507 independent SNPs genome-wide. The tree presented at the top right of the panel represents the branching model for these populations[27-30]. (**c**) Schematic representation of the different population comparisons, indicated by arrows, used for the detection of differentially methylated sites (DMS) between groups.

Figure 2 | DNA methylation profiles and functional differentially methylated regions. (**a-d**) PCA of genome-wide DNA methylation profiles for the different population comparisons. (**e-f**) Gene ontology (GO) enrichment analysis for (**e**) *recent* DMS and (**f**) *historical* DMS. The top 15 and 5 GO categories for biological processes and molecular functions, respectively, are shown, together with the log-transformed FDR-adjusted enrichment *P*-values. The complete lists of enriched categories (FDR adjusted $P < 0.05$) for each dataset are presented in Supplementary Tables 2 and 3.

Figure 3 | Contribution of genetic variation to the DNA methylation levels. (**a**) Proportion of methylation sites that are associated with a nearby genetic variant (in grey) and among different DMS sets (in colour). The numbers in the bars correspond to the total number of DMS per population comparison. *P*-values were calculated by resampling. (**b**) Proportion of the variance of DNA methylation explained by nearby genetic variants (R^2) for the various meQTL sets, in each population. The *P*-values (Mann-Whitney U-test) obtained indicate a significant skew in the R^2 distribution of the various meQTL-DMS sets (in color) with respect to that of all meQTLs (in grey) in the corresponding population. R^2 values are

higher for meQTLs associated with *historical* DMS (11.5% [10.7%-12.3%] and 10.0% [8.9%-11.2%] in w-RHG and f-AGR respectively) than for those related to *recent* DMS (6.5% [5.7%-7.2%] and 6.8% [6.1%-7.4%] in w-AGR and f-AGR respectively). NS not significant, *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$. (**c-e**) meQTLs detected in all populations but presenting different allelic frequencies in RHGs and AGRs. The RHG-AGR mean F_{ST} values for the SNPs concerned were higher (0.15, 0.19 and 0.10, respectively) than that observed genome-wide ($F_{ST} < 0.03$). (**f**) Population-specific meQTL. The SNP rs262643, which is present at similar frequencies in all populations, is associated with methylation differences at *PRKCZ* only in RHGs. (**c-f**) The three plots on the left represent the distribution of M-values as a function of genotype. The minor allele frequency of each meQTL is presented for each population. Red lines indicate the fitted linear regression model for M-value ~ genotype for each population. The forest plots on the right represent the estimated β, corresponding to the slope of the linear regression.

Figure 4 | Selection signals at genetic variants associated to DNA methylation levels. (**a**) Odds ratios measuring the enrichment in high F_{ST} values among meQTLs, with respect to the remainder of the genome, in the different population comparisons. For each comparison, the 95th percentile of the F_{ST} distribution of all SNPs was calculated: 0.11 for w-RHG *vs*. w-AGR, 0.15 for e-RHG *vs*. e-AGR, 0.03 for w-AGR *vs*. f-AGR, and 0.11 for w-RHG *vs*. f-AGR. *P*-values were calculated using a Cochran–Mantel–Haenszel test. (**b**) Distributions of |iHS| values for the different meQTL datasets (in colour) as compared to the remainder of the genome (in grey). *P*-values were estimated using a Mann-Whitney U-test. For both F_{ST} and |iHS|, we considered only SNPs with an LD $r^2 < 0.8$. NS not significant, ***$P < 0.001$.

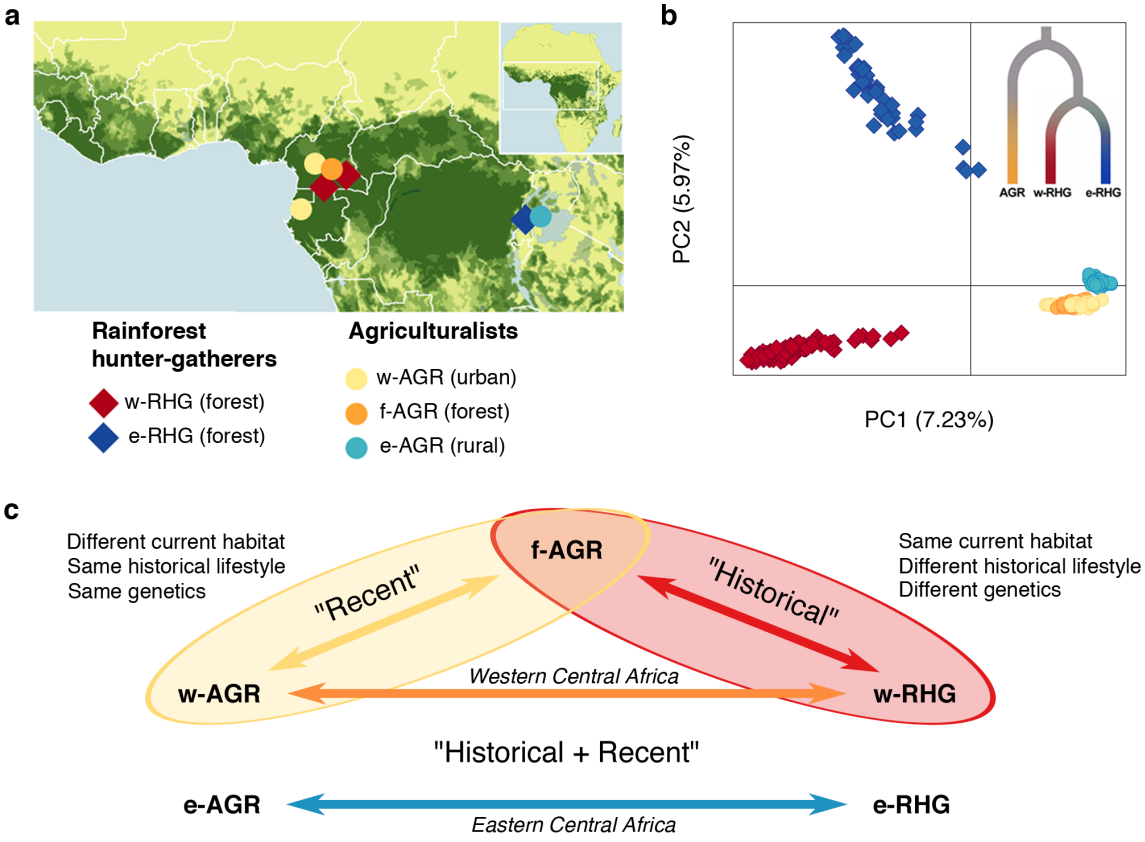

a
Rainforest hunter-gatherers
w-RHG (forest)
e-RHG (forest)
Agriculturalists
w-AGR (urban)
f-AGR (forest)
e-AGR (rural)
b
PC2 (5.97%)
PC1 (7.23%)
AGR
w-RHG
e-RHG
c
Different current habitat
Same historical lifestyle
Same genetics
"Recent"
f-AGR
"Historical"
Same current habitat
Different historical lifestyle
Different genetics
w-AGR
Western Central Africa
w-RHG
"Historical + Recent"
e-AGR
Eastern Central Africa
e-RHG

Figure 1

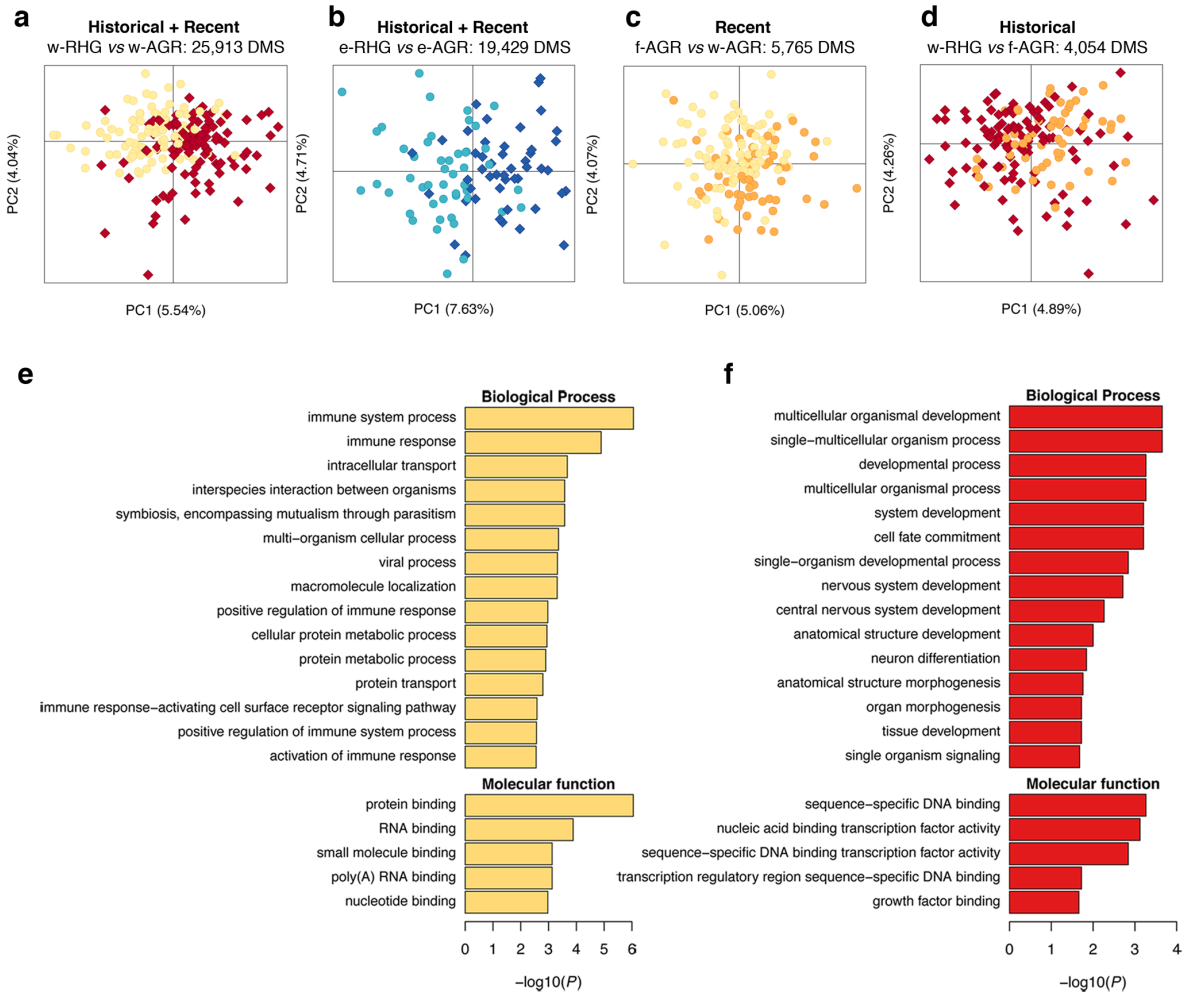

a
Historical + Recent
w-RHG vs w-AGR: 25,913 DMS
PC2 (4.04%)
PC2 (4.71%)
PC1 (5.54%)
b
Historical + Recent
e-RHG vs e-AGR: 19,429 DMS
PC2 (4.07%)
PC1 (7.63%)
c
Recent
f-AGR vs w-AGR: 5,765 DMS
PC2 (4.26%)
PC1 (5.06%)
d
Historical
w-RHG vs f-AGR: 4,054 DMS
PC1 (4.89%)
e
Biological Process
immune system process
immune response
intracellular transport
interspecies interaction between organisms
symbiosis, encompassing mutualism through parasitism
multi−organism cellular process
viral process
macromolecule localization
positive regulation of immune response
cellular protein metabolic process
protein metabolic process
protein transport
immune response−activating cell surface receptor signaling pathway
positive regulation of immune system process
activation of immune response
Molecular function
protein binding
RNA binding
small molecule binding
poly(A) RNA binding
nucleotide binding
0 1 2 3 4 5 6
−log10(P)
f
Biological Process
multicellular organismal development
single−multicellular organism process
developmental process
multicellular organismal process
system development
cell fate commitment
single−organism developmental process
nervous system development
central nervous system development
anatomical structure development
neuron differentiation
anatomical structure morphogenesis
organ morphogenesis
tissue development
single organism signaling
Molecular function
sequence−specific DNA binding
nucleic acid binding transcription factor activity
sequence−specific DNA binding transcription factor activity
transcription regulatory region sequence−specific DNA binding
growth factor binding
0 1 2 3 4
−log10(P)

Figure 2

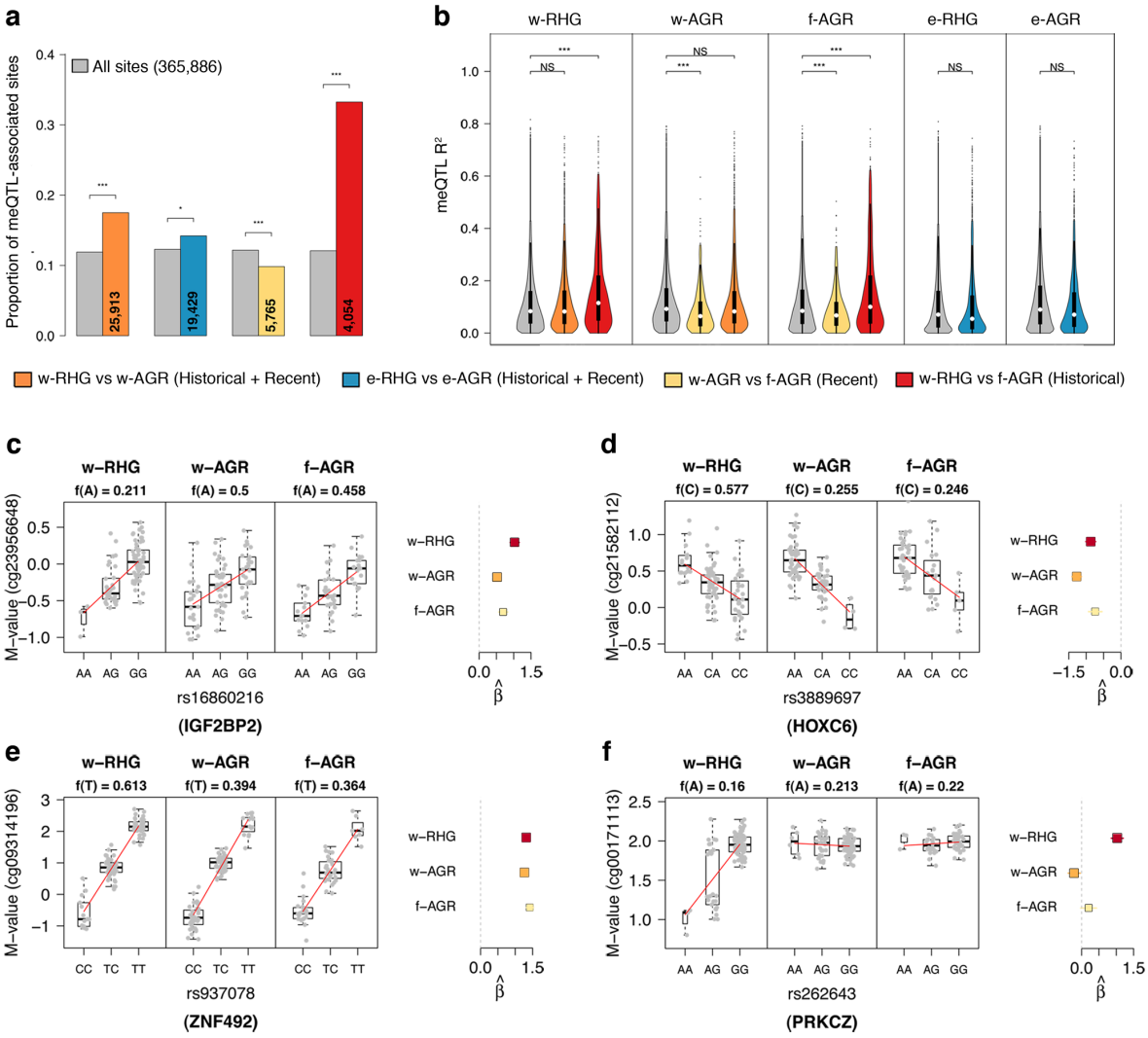

a
Proportion of meQTL-associated sites
All sites (365,886)
25,913
19,429
5,765
4,054
b
w-RHG
w-AGR
f-AGR
e-RHG
e-AGR
meQTL R²
NS
w-RHG vs w-AGR (Historical + Recent)
e-RHG vs e-AGR (Historical + Recent)
w-AGR vs f-AGR (Recent)
w-RHG vs f-AGR (Historical)
c
w–RHG
w–AGR
f–AGR
f(A) = 0.211
f(A) = 0.5
f(A) = 0.458
M–value (cg23956648)
AA AG GG
rs16860216
(IGF2BP2)
d
f(C) = 0.577
f(C) = 0.255
f(C) = 0.246
M–value (cg21582112)
AA CA CC
rs3889697
(HOXC6)
e
f(T) = 0.613
f(T) = 0.394
f(T) = 0.364
M–value (cg09314196)
CC TC TT
rs937078
(ZNF492)
f
f(A) = 0.16
f(A) = 0.213
f(A) = 0.22
M–value (cg00171113)
rs262643
(PRKCZ)

Figure 3

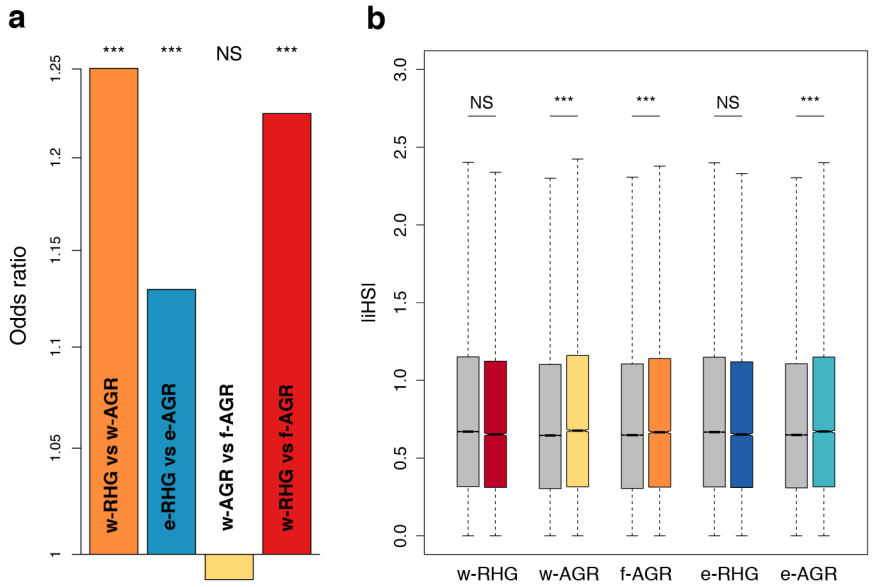

a
b
Odds ratio
1
1.05
1.1
1.15
1.2
1.25
***
***
NS
***
w-RHG vs w-AGR
e-RHG vs e-AGR
w-AGR vs f-AGR
w-RHG vs f-AGR
liHSI
0.0
0.5
1.0
1.5
2.0
2.5
3.0
NS
***
***
NS
***
w-RHG
w-AGR
f-AGR
e-RHG
e-AGR

Figure 4

Table 1. Sampling location, historical modes of subsistence and current habitat of populations in the study

Group	Population	Sampling location(s)	Historical mode of subsistence	Current habitat/lifestyle	N*	N†	N§
w-RHG	Baka	Lomié-Messok, Salapoumbe, Oveng-Djoum, Southeast Cameroon	Hunter-gatherers	Villages in the equatorial rainforest. Slash-and-burn agriculture, subsistence farming, hunting and gathering in the equatorial forest	78	73	68
w-RHG	Baka	Minvoul, Northeast Gabon	Hunter-gatherers	Villages in the equatorial rainforest. Slash-and-burn agriculture, subsistence farming, hunting and gathering in the equatorial forest	34	30	29
e-RHG	Batwa	Southwest Uganda	Hunter-gatherers¶	Villages near the forest after the creation of the Bwindi Impenetrable Forest National Park. Subsistence farming, hunting and gathering in the equatorial forest before settling	47	47	47
w-AGR	Nzebi	Libreville, Gabon	Agriculturalists	Urban	55	55	55
w-AGR	Fang#	Yaoundé, Cameroon	Agriculturalists	Urban	39	39	39
e-AGR	Bakiga	Southwest Uganda	Agriculturalists	Villages in rural, deforested areas. Subsistence farming in stable deforested area.	48	48	48
f-AGR	Nzime	Lomié-Messok, Southeast Cameroon	Agriculturalists	Villages in the equatorial rainforest, shared habitat with w-RHG Baka from Cameroon (mostly from the Lomié region). Slash-and-burn agriculture, forest hunting	61	60	59

*Sample sizes before normalization and filtering; †Sample sizes, after normalization and filtering, used for methylation analyses; §Sample sizes, after SNP imputation and filtering for low call rates, used for meQTL mapping; ¶Although, at present, the Batwa RHG do not live in the forest, they hunted and gathered in the Bwindi Impenetrable Forest in southwest Uganda until it became a national park in 1991. All individuals included in this study were born and raised in the equatorial forest, where they lived in non-permanent camps; #This sample corresponds to a composite sample of Bantu-speaking individuals from Yaoundé, mostly belonging to the Fang ethnic group.

7.3 Conclusions et discussion

7.3.1 Résumé des résultats et nouveautés

L'étude des variations des profils de méthylation de l'ADN entre plusieurs populations de RHG et d'AGR d'Afrique centrale constitue la première étude de l'effet respectif des facteurs environnementaux et génétiques sur la variabilité du niveau de méthylation de l'ADN entre populations humaines sur des tissus primaires et à l'échelle du génome. En effet, l'immense majorité des études de la variation des profils de méthylation entre populations ont été réalisées sur des lignées cellulaires, en comparant des populations urbaines issues de continents différents, maximisant ainsi la différenciation génétique entre les populations et minimisant les différences environnementales (Fraser et al. 2012, Heyn et al. 2013, Moen et al. 2013). Notre étude a permis de mettre en évidence que les différences d'habitat récent, d'une part, et les différences d'environnement et d'histoire génétique anciennes, d'autre part, entraînent des variations de niveau de méthylation à plusieurs milliers de sites. De façon intéressante, ces deux jeux de DMS (sites différentiellement méthylés) ne se chevauchent pas, et affectent des gènes ayant des fonctions biologiques différentes. Les DMS entre AGR vivant dans des milieux urbains et des milieux forestiers, dits DMS récents, sont ainsi localisés dans des gènes impliqués dans l'immunité alors que les DMS entre AGR forestiers et RHG, dits DMS historiques, se trouvent dans des gènes en rapport avec les processus de développement.

Nous avons également détecté des associations significatives entre variations génétiques et variations du niveau de méthylation (meQTL) à environ 12% des sites mesurés. Parmi ces sites, 90% présentent une association avec un variant génétique dans l'ensemble des populations étudiées, ce qui est en contradiction avec des résultats précédents (Heyn et al. 2013, Moen et al. 2013). Cela pourrait être dû au fait que, contrairement à ces deux études, nous analysons des données de méthylation provenant d'un tissu complexe constitué de nombreux types cellulaires, mais également à un manque de puissance pour détecter des meQTL spécifiques d'une population. Les SNP sont enrichis en signatures de sélection positive récente. En particulier, ils présentent un excès significatif de signature de sélection récente (<30 000 ans) chez les Agriculteurs.

Enfin, en croisant les listes de DMS et de sites associés aux meQTLs, nous

avons montré que les DMS historiques, au contraire des DMS récents, sont enrichis en associations avec des meQTL, et que la proportion de variance expliquée par le SNP est plus élevée que pour le reste des associations. Nous avons donc mis en évidence l'existence de deux groupes de DMS indépendants : ceux associés à des différences d'habitat récent (ville *vs.* forêt), localisés dans des gènes liés à l'immunité et ne présentant pas d'association avec des facteurs génétiques, et ceux associés à des différences de profil génétique et d'environnement passé (AGR *vs.* RHG Pygmées), localisés dans des gènes du développement et largement expliqués par la différenciation entre les populations des variants génétiques sous-jacents.

7.3.2 Intérêts

Notre étude permet, pour la première fois, d'évaluer les différences de méthylation de l'ADN mesurées sur tissu primaire, le sang complet, entre des populations d'agriculteurs vivant dans des habitats différents, et semblables par ailleurs. Le fait que des différences d'habitat récent corrèlent avec des modifications de méthylation de l'ADN dans des gènes de l'immunité suggère un lien fonctionnel entre les différences du paysage épigénétique et les réponses de l'hôte à des expositions à différents pathogènes dues aux différences d'habitats. Une étude de Idaghdour et al. (2008) a relevé une différence de niveau d'expression de gènes reliés à certaines fonctions immunitaires sur un modèle similaire de populations de nomades des montagnes et de sédentaires urbains du Maroc, et s'est brièvement intéressé aux variations de la méthylation de l'ADN, sans trouver de différences entre les populations, probablement à cause de la faible taille d'échantillon (<20 individus par échantillon), et du nombre restreint de sites mesurés (1 500 sites répartis sur 800 gènes). Une étude de l'expression des gènes nous permettrait d'obtenir plus d'informations sur le lien entre environnement, variation des profils de méthylation et phénotype.

Nous avons également montré que les facteurs génétiques et environnementaux ont tous deux un impact non négligeable sur la variation des profils de méthylation. En particulier, l'absence d'association entre DMS récents et facteurs génétiques suggèrent que les populations pourraient répondre immédiatement aux changements environnementaux via des modifications de leur paysage épigénétique de façon indépendante des mutations génétiques. Nous avons vu dans l'introduction (chapitre 5) qu'une modification du profil de méthylation due à des facteurs

environnementaux n'est que très rarement héritable (Heard and Martienssen 2014). Les DMS récents sont probablement expliqués par le fait que dans chaque population, des profils de régulation de l'expression des gènes spécifiques sont établis au cours de la vie, en réponse aux caractéristiques de chaque habitat (dont en particulier l'exposition aux pathogènes), de façon similaire chez les différents individus au sein d'une population.

Enfin, nos travaux indiquent que les meQTL ont été des cibles de la sélection positive, suggérant qu'elle a agi, directement ou indirectement, sur les variations du profil de méthylation. Ces résultats pris tous ensemble ouvrent ainsi une réflexion sur le rôle de l'épigénétique dans l'adaptation des populations à leur environnement sur différentes échelles de temps.

DISCUSSION

Chapitre 8

Perspectives

8.1 Vers un tableau plus complet de l'action de la sélection positive sur le génome humain

8.1.1 Au-delà des gènes : quel impact de la sélection positive et régions régulatrices du génome ?

Nos travaux sur les balayages sélectifs récents ont permis de mettre en évidence leur rôle non négligeable dans l'évolution du génome humain, puisque nous estimons à une petite centaine le nombre de gènes ayant évolués selon ce modèle dans chacune des trois populations étudiées. Cependant, ce chiffre pourrait être sous-estimé, étant donné qu'il ne prend en compte que les régions géniques, soit environ 40% du génome. Or, parmi les signaux de sélection détectés dans notre étude, beaucoup sont situés en dehors des régions géniques.

L'hypothèse que les régions régulatrices pourraient avoir joué un rôle majeur dans l'adaptation humaine remonte à un article de King and Wilson en 1975 et a été développée dans plusieurs publications récentes (Akey 2009, Wray 2007, Yue et al. 2014). Elles seraient ainsi la cible d'événements de sélection positive environ 10x plus fréquemment que les mutations non synonymes (Fraser 2013). Si la mutation entraînant la persistance de la production de lactase à l'âge adulte, déjà abordée dans l'introduction, est l'exemple emblématique de tels phénomènes, beaucoup d'autres mutations affectant l'expression de gènes pourraient être sous sélection. Deux études ont conclu par exemple que la majorité des balayages sélectifs touchent des régions régulatrices (Enard et al. 2014, Grossman et al. 2013). Divers travaux montrent également que les mutations associées à des variations du niveau d'expression

des gènes sont enrichies en signaux de sélection positive (Kudaravalli et al. 2009, Torgerson et al. 2009, Ye et al. 2013), soulignant l'importance de s'intéresser à ces régions pour mieux comprendre l'histoire de l'adaptation des populations humaines à leur environnement.

L'un des problèmes majeurs lié à l'étude du rôle des régions régulatrices dans l'histoire évolutive humaine reste l'identification de la mutation ciblée par la sélection et du phénotype en résultant. Ces dernières années, de nombreux efforts ont été réalisés pour décrire et cartographier les régions régulatrices du génome. Ainsi, l'analyse de différentes données épigénétiques et d'expression des gènes ont permis d'identifier un certain nombre de promoteurs et d'enhancers (ENCODE Project Consortium 2012, Ernst et al. 2011, Heintzman et al. 2007, Kim et al. 2005). Nos travaux ont permis d'établir une liste de régions non codantes candidates et constituent une base pour de futures études visant à mieux caractériser les événements de sélection ciblant des régions régulatrices du génome. Une première étape consisterait à identifier plus précisément la mutation ciblée par la sélection en utilisant des tests spécialement conçus à cet effet (Grossman et al. 2010). Après identification de l'élément régulateur concerné à l'aide des cartes épigénétiques, le croisement de ces résultats avec les résultats de GWAS et des données fonctionnelles dans le tissu pertinent (trancriptome, métabolome, protéome) permettrait d'identifier plus précisément la fonction biologique visée et éventuellement la pression de sélection (Akey 2009, Lachance and Tishkoff 2013, Scheinfeldt and Tishkoff 2013, Schraiber et al. 2013, Vernot et al. 2012, Ward and Kellis 2012b). Cela permettrait d'obtenir une image plus complète de l'effet de la sélection positive sur l'évolution du génome humain.

8.1.2 Les autres modes de sélection positive : quel impact sur la diversité phénotypique humaine ?

Notre étude ainsi que les nombreux travaux visant à répertorier les traces de sélection positive sur le génome humain se sont concentrées sur la détection des balayages sélectifs. En effet, les statistiques utilisées dans la première partie de cette thèse permettent de détecter les signatures des événements de sélection à fort coefficient touchant des allèles dont la fréquence dans la population avant le début de l'événement de sélection était faible. Si les balayages sélectifs plus anciens

ont été étudiés en utilisant d'autre méthodes (Bustamante et al. 2005, Chimpanzee Sequencing and Analysis Consortium 2005, Clark et al. 2003, Diller et al. 2002, Nielsen et al. 2005a), une évaluation complète de l'importance de la sélection positive dans l'évolution humaine nécessiterait également de s'intéresser au autres modes de sélection positive : la sélection à plus faible coefficient, la sélection sur allèles pré-existants à fréquence modéré et la sélection polygénique (Fu and Akey 2013, Hermisson and Pennings 2005, Pritchard and Di Rienzo 2010, Scheinfeldt and Tishkoff 2013, Teshima et al. 2006).

Cependant, ces modes de sélection sont difficiles à détecter en utilisant les outils classiques (Chevin and Hospital 2008, Hancock et al. 2011, Prezeworski et al. 2005, Pritchard et al. 2010, Teshima and Przeworski 2006, Teshima et al. 2006), car ils ne produisent pas les mêmes signatures moléculaires que les balayages sélectifs (cf. 2 et figure refregimesC et D), principalement en ayant un effet local plus faible sur la diversité génétique et le déséquilibre de liaison. Si la sélection locale d'un trait monogénique à partir d'une mutation déjà présente dans la population peut être détectée en utilisant des tests basés sur la différenciation de populations ou des test composites (Grossman et al. 2010, Innan and Kim 2008, Prezeworski et al. 2005), un des enjeux majeurs de ces prochaines années est la détection de la sélection polygénique, puisqu'elle pourrait représenter un mode d'adaptation majeur à l'environnement (Hernandez et al. 2011, Messer and Petrov 2013b).

En effet, de nombreux traits comme la taille ou la résistance aux pathogènes sont déterminés non par un seul mais par plusieurs gènes, et la sélection polygénique permet une modulation fine de ces phénotypes dans le cadre d'une adaptation à un changement environnemental. C'est ainsi le cas de la petite taille des populations de chasseurs-cueilleurs d'Afrique Centrale, les Pygmées, qui est le résultat d'une adaptation polygénique (Bryc et al. 2010, Jarvis et al. 2012, Lachance et al. 2012, Verdu et al. 2013). De façon notable, des adaptations convergentes ont été trouvées dans des populations de chasseurs-cueilleurs d'Asie du Sud-Est et d'Océanie, laissant penser que le phénotype pygmée est le résultat d'une adaptation à la forêt tropicale (Perry and Dominy 2009, Perry et al. 2014). Plusieurs méthodes ont été proposées afin de détecter de tels événements : rechercher les enrichissements en signaux de sélection positives de gènes appartenant à une même voie métabolique ou régulatrice (Berg and Coop 2014, Daub et al. 2013, Fraser 2013, Simonson et al. 2010), ou bien

de SNP associés au même phénotype (Turchin et al. 2012, Zhang et al. 2013), ou enfin corréler des fréquences alléliques à plusieurs SNP avec des facteurs environnementaux (Fumagalli et al. 2011, Hancock et al. 2011).

Enfin, l'étude des génomes d'Hommes archaïques laissent penser que l'introgression adaptative pourrait également avoir eu un impact sur l'évolution des populations humaines. En effet, si certaines régions du génome humain, notamment en lien avec la reproduction, semblent avoir été purgées des mutations provenant de Néandertal ou Denisova, les gènes en rapport avec des fonctions telles que l'immunité semblent avoir au contraire particulièrement conservé ces mutations, et montrent des traces de sélection positive (Abi-Rached et al. 2011, Green et al. 2010, Ségurel and Quintana-Murci 2014, Vernot and Akey 2014). C'est ainsi le cas du gène de l'immunité innée *STAT2* sous sélection en Papouasie Nouvelle-Guinée et du cluster *OAS* en Mélanésie (Mendez et al. 2012, 2013). D'autres phénotypes semblent avoir bénéficié de l'introgression de Denisova, comme par exemple l'adaptation à la haute altitude au Tibet (Huerta-Sánchez et al. 2014), ou de Néandertal, comme l'adaptation à l'exposition aux UV en Asie de l'Est (Ding et al. 2014) et le catabolisme des lipides en Europe (Khrameeva et al. 2014). Cela suggère qu'une partie de la variabilité phénotypique humaine pourrait provenir de tels événements. Ainsi, les études en cours et à venir de l'impact des modes alternatifs de sélection positive différents ainsi que de l'introgression adaptative de portions de génomes d'Hommes archaïques sur le génome humain sont très prometteuses. Elles pourraient permettre d'avoir une vision beaucoup plus complète du rôle de l'environnement dans l'évolution de la diversité génétique et phénotypique humaine.

8.2 Reproductibilité et effets phénotypiques des variations épigénétiques associées à l'environnement

8.2.1 Les variations de méthylation liées à l'environnement : quel impact sur l'expression des gènes ?

Nous avons mis en évidence chez des populations d'Afrique Centrale l'effet des changements d'habitat récent sur le niveau de méthylation de certaines régions

génomiques, en particulier des gènes liés à l'immunité. Cette observation soulève la question de l'impact phénotypique de telles modifications épigénétiques. Il serait par exemple intéressant de rechercher si les gènes de l'immunité qui montrent des variations de niveau de méthylation de leurs promoteurs entre populations montrent également des changements d'expression. Plus généralement, corréler des données d'expression avec des données de méthylation pourrait permettre d'évaluer plus concrètement si des différences épigénétiques reliées à des différences environnementales sont effectivement corrélées à des variations du transcriptome, et constitue la suite logique de cette étude. Cela permettrait d'apporter des informations sur d'éventuels effets phénotypiques dus à des variations environnementales et liés à des variations épigénétiques.

Cependant, si on observe une corrélation négative entre le niveau d'expression et de méthylation des promoteurs entre différents types cellulaires pour un même gène et entre gènes au sein d'un même type cellulaire (Bell et al. 2011, Jones 2012), cette corrélation est beaucoup moins évidente entre individus pour un même type cellulaire et un même gène. Différentes études ont en effet relevé des corrélations aussi bien positives que négatives, et souvent relativement faibles, entre ces deux paramètres chez différents individus (Gutierrez-Arcelus et al. 2013, Heyn et al. 2013, Lam et al. 2012). Gutierrez-Arcelus et al. suggèrent que des mécanismes partiellement indépendants sont à l'origine des associations entre niveau de transcription et de méthylation des promoteurs observées au cours du développement ou de la différenciation cellulaire, d'une part, et entre individus, d'autre part. Si les mécanismes moléculaires de la différenciation cellulaire et du développement sont largement étudiés, ceux qui sont en jeu dans les modifications épigénétiques par des facteurs environnementaux commencent tout juste à être appréhendés. Zhao et al., par exemple, ont montré un lien entre exposition aux quinones à effet rédox et l'augmentation du niveau de fer labile dans la cellule. Cette modification du niveau de fer utilisable entraîne une augmentation de l'activité d'une enzyme de la famille *TET*, dont le fer II est un co-facteur, et promeut par conséquent la transformation de 5-mC en 5-hmC qui constitue la première étape du processus déméthylation. Il est crucial dans les années à venir de se pencher sur l'étude *in vitro* des effets de l'exposition de cellules à certains facteurs environnementaux sur leur équilibre chimique, leur épigénome, leur transcriptome et leur métabolome, et de confirmer ces résultats *in*

vivo, afin d'explorer leur potentielle diversité.

Mieux comprendre l'effet de ces modifications de niveau de méthylation sur les phénotypes apparaît également nécessaire (Szyf 2011). Des projets comme ENCODE (ENCODE Project Consortium 2004, 2012, Kellis et al. 2014), et The Epigenome Roadmap (Bernstein et al. 2010, Roadmap Epigenomics Consortium et al. 2015) ont commencé récemment à fournir de nombreuses informations sur l'accessibilité de l'ADN aux enzymes de restrictions, sur l'état de différents acteurs épigénétiques, sur les sites de fixations utilisés par des facteurs de transcription et sur le transcriptome dans plusieurs types cellulaires issus de tissus sains ou malades. L'intégration de l'ensemble de ces données permet de mieux appréhender les différents mécanismes en jeu lors du développement et de la différenciation cellulaire, mais également de définir les éléments fonctionnels du génome ainsi que les bases moléculaires de certains traits phénotypiques et maladies. Une comparaison de l'état de diverses marques épigénétiques au niveau des éléments régulateurs du génomes entre différents individus pour le même tissu, couplée à des données d'expression des gènes pourrait permettre une meilleure compréhension du lien entre variations épigénétiques et phénotypiques et ainsi apporter un éclairage sur l'impact phénotypique de modifications environnementales.

8.2.2 Action de l'environnement sur l'épigénome : les mêmes causes produisent-elles les mêmes effets ?

Dans cette thèse, nous avons montré que 57% des sites différentiellement méthylés entre agriculteurs urbains et agriculteurs forestiers sont également différentiellement méthylés dans la même direction entre agriculteurs urbains et chasseurs-cueilleurs forestiers, ce qui représente un chevauchement très significatif. Ces DMS communs aux deux groupes constituent donc un ensemble de sites dont le niveau de méthylation est associé à l'habitat (urbain ou forestier) des populations et indépendant des différences génétiques et environnementales passées pouvant les séparer. Ces résultats suggèrent un effet au moins partiellement identique de l'environnement forestier sur des populations présentant des profils génétiques différents (agriculteurs et chasseurs-cueilleurs), posant la double question de la spécificité et de la reproductibilité de l'effet des paramètres environnementaux sur l'épigénome. Des travaux chez la souris ont montré que l'exposition à des plastiques, des pesticides, des dioxines ou des

carburants provoquent des effets différents, spécifiques de chaque composant, sur le méthylome des individus (Manikkam et al. 2012). Les mécanismes moléculaires de cette spécificité doivent encore être explorés. Au contraire, ces travaux révèlent que l'exposition à un même composant entraîne des modifications très semblables chez plusieurs groupes de souris (Manikkam et al. 2012), montrant une reproductibilité des effets de l'environnement sur des individus génétiquement similaires. Cependant, ce modèle ne permet pas d'étudier si des populations génétiquement différentes exposées à un même environnement vont présenter les mêmes différences en terme de profil épigénétique.

Différents travaux ont déjà montré que des paramètres environnementaux tels que le niveau socio-économique ou l'exposition précoce à la fumée de cigarette avait un impact similaire sur le profil de méthylation, quel que soit le profil génétique des individus (Drake et al. 2015, Lam et al. 2012). Notre étude permet d'apporter de nouveaux indices en faveur de la reproductibilité spatiale des effets de l'environnement sur le méthylome. La suite logique, afin de confirmer cette hypothèse serait de déterminer si des populations Pygmées vivant en ville présenteraient des profils de méthylation similaires à ceux des agriculteurs urbains pour les DMS associés à l'habitat. Cependant, on trouve très peu de Pygmées Baka nés et élevés en milieu urbain, et il est donc impossible de « fermer la boucle ». Pour mieux étudier cette question, nous proposons un nouveau projet visant à comparer les profils de méthylation de plusieurs populations vivant soit en milieu urbain, soit en milieu rural. En effet, dans les grandes villes cosmopolites, des individus issus de plusieurs populations cohabitent dans le même environnement. Le but de cette étude serait de comparer le profil de méthylation de plusieurs populations, par exemple européennes, asiatiques, africaines du Nord et d'Afrique sub-saharienne vivant dans une même ville aux profils de méthylation d'individus de ces mêmes populations vivant en milieu rural. Cela permettrait de déterminer si le profil de méthylation de ces différentes populations est affecté de la même façon par les différences environnementales entre milieu urbain et rural. Une comparaison des profils de méthylation des populations européennes urbaines et rurales avec celui d'une population européenne vivant à la montagne permettrait ensuite d'identifier les sites présentant une méthylation spécifique du milieu urbain.

La reproductibilité temporelle de ces différences de méthylation au cours du temps

est également intéressante. En effet, s'il est établi que les modifications des profils de méthylation de l'ADN par l'environnement n'est que très rarement transmissible à la génération suivante (cf. chapitre 5), il est important de se demander si des populations confrontées aux mêmes environnements à différentes époques montrent des profils de méthylation similaires. Une étude récente a montré qu'il était possible d'accéder au profil de méthylation d'un homme moderne ayant vécu il y a environ 4 000 ans et donc d'accéder à des informations concernant l'état chromatinien et la régulation des gènes à partir de données d'ADN ancien (Pedersen et al. 2014). Etablir de tels profils pour différents individus à partir d'ADN anciens pourrait ainsi permettre d'évaluer si les effets de l'environnement sur le méthylome sont constants au cours du temps et d'évaluer les effets des changements environnementaux sur le profil épigénétique d'une population au cours des générations (en comparant par exemple le profil de méthylation de l'ADN d'individus ayant vécu à différentes périodes avant et après la révolution industrielle en Europe). Etudier l'indépendance par rapport au profil génétique et la reproductibilité temporelle des effets de l'environnement sur l'état épigénétique d'un tissu humain serait un premier pas dans l'étude du rôle de l'épigénétique dans l'adaptation rapide des populations à leur environnement.

8.3 L'environnement et la diversité génétique, épigénétique et phénotypique : un modèle d'adaptation plus complexe ?

Nos travaux montrent enfin que les mutations génétiques expliquant une partie de la variabilité des profils de méthylation de l'ADN entre individus portent des traces de sélection positive population-spécifique relativement récente, particulièrement chez les agriculteurs. Sans préjuger du lien direct ou indirect entre variations génétiques et épigénétiques, ce résultat suggère que la sélection a ciblé particulièrement des mutations favorisant la fixation d'un état épigénétique donné. A cela s'ajoutent nos conclusions sur l'existence d'un effet spécifique de l'environnement forestier sur le méthylome indépendamment du profil génétique des populations. Considérés ensembles, ces résultats suggèrent un rôle pour l'épigénétique dans l'adaptation à l'environnement. Nous proposons donc un modèle d'adaptation des populations humaines à l'environnement à différentes échelles de temps intégrant la génétique

et l'épigénétique (figure 13). Il est important de noter qu'il ne s'agit pas ici de réhabiliter la théorie lamarckienne de l'évolution impliquant l'héritabilité des caractères acquis. En effet, de plus en plus d'études montrent l'extrême rareté de la transmission de modifications épigénétiques d'origine environnementale à travers les générations (cf. chapitre 5). C'est pourquoi notre modèle ne considère pas la possibilité d'une héritabilité des modifications épigénétiques, contrairement à un modèle par ailleurs semblable proposé par Klironomos et al. (2013). Il s'agit plutôt de souligner le rôle potentiel de l'épigénétique comme facteur d'adaptation à court terme à l'environnement.

Dans notre modèle, une population soumise à un changement environnemental (passage de l'environnement 1 à 2) pourrait réagir à court terme en modifiant l'expression de certains gènes associé à une modification de l'état épigénétique de leurs régions régulatrices, par exemple en provoquant leur passage dans l'hétérochromatine. Si les nouvelles conditions environnementales sont maintenues au cours du temps, elles peuvent reproduire ces effets sur les générations successives. Si, à un moment donné de l'évolution de cette population, il se produit par hasard une mutation dans un élément régulateur permettant de fixer l'état de régulation de l'expression du gène favorable dans l'environnement 2, ce phénotype pourra alors être transmis à la descendance selon les lois de la génétique mendélienne classique. Une adaptation à long terme de la population à cet environnement est alors possible via un phénomène de sélection darwinienne classique. Le fait qu'une proportion importante du méthylome soit associé avec des facteurs génétiques lui conférant une héritabilité à travers les générations (Mendizabal et al. 2014) ainsi que l'existence de nombreux événements de sélection touchant des régions régulatrices du génome (Enard et al. 2014, Grossman et al. 2013) plaident en faveur d'un tel modèle.

L'un des avantages de fixer l'état de régulation du gène de façon génétique est de permettre une relaxation de la pression environnementale sur le profil de méthylation de l'élément régulateur, qui pourra alors « dériver ». Un autre modèle d'adaptation par accommodation génétique a été proposé récemment (Crispo 2007, Feinberg and Irizarry 2010, Xu et al. 2015), qui suggère que la sélection favoriserait les SNPs associés avec une plus grande plasticité de l'état épigénétique et de la régulation de l'expression des gènes (Feinberg and Irizarry 2010, Xu et al. 2015). En effet, de telles mutations favoriseraient une réponse rapide à des changements

environnementaux. Ces deux modèles ne sont pas incompatibles, le nôtre supposant une adaptation stable à un changement environnemental à long terme, alors que celui de le modèle de l'accommodation permet une adaptation rapide lorsque les populations sont confrontées à une grande variabilité environnementale.

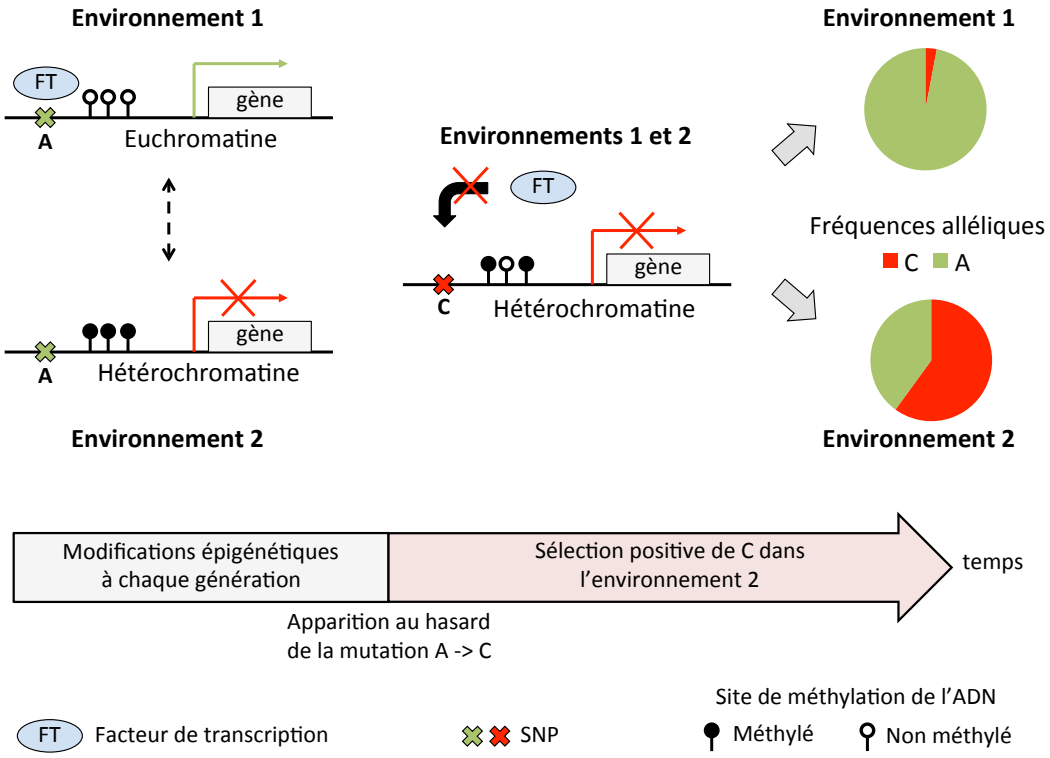


Fig. 13 L'adaptation des populations humaines à leur environnement : un modèle intégrant la génétique et l'épigénétique.

La validation de ces modèles suppose dans un premier temps de croiser les données existantes sur les régions sous sélection du génome avec les cartes génomiques d'éléments régulateurs afin d'identifier des mutations candidates. Une étude de leurs effets phénotypiques sur l'expression des gènes et de leur corrélation avec des modifications locales de l'état de la chromatine permettrait ensuite de disséquer les mécanismes d'adaptation par régulation de l'expression des gènes. Il serait ainsi possible d'évaluer la pertinence biologique de chacun de ces modèles, et ainsi de mieux apprécier les rôles respectifs joués par la génétique et l'épigénétique dans l'adaptation à l'environnement.

Chapitre 9

Conclusion générale

Les résultats présentés ici contribuent à une meilleure compréhension de l'impact de l'environnement sur la diversité génétique et épigénétique humaine. Ces travaux ont permis de tirer un certain nombre de conclusions. Sur le plan méthodologique d'abord, ils soulignent l'importance de vérifier l'adéquation des méthodes utilisées avec les données à analyser, via le développement de méthodes insensibles aux caractéristiques des données de séquençage à haut débit pour réaliser des études de génétique des populations. Sur le plan des interactions entre environnement et diversité génétique ensuite, ils montrent que l'étude des données récentes de séquençage à faible profondeur nous permettent de retrouver des signaux bien connus de balayage sélectif, et que des classes de mutations fonctionnelles sont enrichies en signaux de sélection, mettant ainsi en évidence que les balayages sélectifs ont bien eu une influence, certes modérée, mais non négligeable sur l'évolution récente du génome humain. Sur le plan des interactions entre environnement, génétique et épigénétique, ils ont permis d'établir que les différences d'habitat récent ou d'environnement ancien et d'histoire génétique entraînent toutes deux des différences de niveau méthylation entre population, mais affectent des sites et des fonctions biologiques très différents, et que 10% des sites de méthylation de l'ADN présentent des variations associées à des mutations génétiques, qui sont elles-mêmes enrichies en signaux de sélection positive. Ces résultats nous poussent à interroger le lien entre variations environnementales, génétiques, épigénétiques et phénotypiques, ainsi que la place de l'épigénétique dans l'adaptation des populations à leur environnement.

BIBLIOGRAPHIE

Bibliographie

Abadie, V., Sollid, L. M., Barreiro, L. B., and Jabri, B. Integration of genetic and immunological insights into a model of celiac disease pathogenesis. *Annual Review of Immunology*, 29 :493–525, 2011. ISSN 1545-3278. doi : 10.1146/annurev-immunol-040210-092915.

Aberg, K. A., McClay, J. L., Nerella, S., Clark, S., Kumar, G., Chen, W., Khachane, A. N., Xie, L., Hudson, A., Gao, G., Harada, A., Hultman, C. M., Sullivan, P. F., Magnusson, P. K. E., and van den Oord, E. J. C. G. Methylome-wide association study of schizophrenia : identifying blood biomarker signatures of environmental insults. *JAMA psychiatry*, 71(3) :255–264, Mar. 2014. ISSN 2168-6238. doi : 10.1001/jamapsychiatry.2013.3730.

Abi-Rached, L., Jobin, M. J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F. A., Kimani, J., Carrington, M., Middleton, D., Rajalingam, R., Beksac, M., Marsh, S. G. E., Maiers, M., Guethlein, L. A., Tavoularis, S., Little, A.-M., Green, R. E., Norman, P. J., and Parham, P. The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science (New York, N.Y.)*, 334(6052) :89–94, Oct. 2011. ISSN 1095-9203. doi : 10.1126/science.1209202.

Aimé, C., Laval, G., Patin, E., Verdu, P., Ségurel, L., Chaix, R., Hegay, T., Quintana-Murci, L., Heyer, E., and Austerlitz, F. Human Genetic Data Reveal Contrasting Demographic Patterns between Sedentary and Nomadic Populations That Predate the Emergence of Farming. *Molecular Biology and Evolution*, 30(12) :2629–2644, Dec. 2013. doi : 10.1093/molbev/mst156. URL http://mbe.oxfordjournals.org/content/30/12/2629.abstract.

Akey, J. M. Constructing genomic maps of positive selection in humans : Where do we go from here ? *Genome Research*, 19(5) :711–722, May 2009. ISSN 1088-9051. doi : 10.1101/gr.086652.108. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.086652.108.

Akey, J. M., Zhang, G., Zhang, K., Jin, L., and Shriver, M. D. Interrogating a High-Density SNP Map for Signatures of Natural Selection. *Genome Research*, 12(12) : 1805 –1814, Dec. 2002. doi : 10.1101/gr.631202. URL http://genome.cshlp.org/content/12/12/1805.abstract.

Alexander, D. H., Novembre, J., and Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9) :1655–1664, Sept. 2009.

ISSN 1088-9051. doi : 10.1101/gr.094052.109. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.094052.109.

Alkorta-Aranburu, G., Beall, C. M., Witonsky, D. B., Gebremedhin, A., Pritchard, J. K., and Di Rienzo, A. The Genetic Architecture of Adaptations to High Altitude in Ethiopia. *PLoS Genetics*, 8(12) :e1003110, Dec. 2012. ISSN 1553-7404. doi : 10.1371/journal.pgen.1003110. URL http://dx.plos.org/10.1371/journal.pgen.1003110.

Allison, A. C. Protection afforded by sickle-cell trait against subtertian malareal infection. *British Medical Journal*, 1(4857) :290–294, Feb. 1954. ISSN 0007-1447.

Allison, A. C. GENETIC FACTORS IN RESISTANCE TO MALARIA. *Annals of the New York Academy of Sciences*, 91(3) :710–729, June 1961. ISSN 00778923. doi : 10.1111/j.1749-6632.1961.tb31102.x. URL http://doi.wiley.com/10.1111/j.1749-6632.1961.tb31102.x.

Amorim, C. E. G., Daub, J. T., Salzano, F. M., Foll, M., and Excoffier, L. Detection of convergent genome-wide signals of adaptation to tropical forests in humans. *PloS One*, 10(4) :e0121557, 2015. ISSN 1932-6203. doi : 10.1371/journal.pone.0121557.

Andersen, K. G., Shylakhter, I., Tabrizi, S., Grossman, S. R., Happi, C. T., and Sabeti, P. C. Genome-wide scans provide evidence for positive selection of genes implicated in Lassa fever. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 367(1590) :868–877, Mar. 2012. ISSN 0962-8436, 1471-2970. doi : 10.1098/rstb.2011.0299. URL http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2011.0299.

Aravin, A. A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K. F., Bestor, T., and Hannon, G. J. A piRNA Pathway Primed by Individual Transposons Is Linked to De Novo DNA Methylation in Mice. *Molecular Cell*, 31(6) :785–799, Sept. 2008. ISSN 10972765. doi : 10.1016/j.molcel.2008.09.003. URL http://linkinghub.elsevier.com/retrieve/pii/S1097276508006199.

Asthana, S., Noble, W. S., Kryukov, G., Grant, C. E., Sunyaev, S., and Stamatoyannopoulos, J. A. Widely distributed noncoding purifying selection in the human genome. *Proceedings of the National Academy of Sciences*, 104(30) :12410–12415, July 2007. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.0705140104. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0705140104.

Ayub, Q., Moutsianas, L., Chen, Y., Panoutsopoulou, K., Colonna, V., Pagani, L., Prokopenko, I., Ritchie, G., Tyler-Smith, C., McCarthy, M., Zeggini, E., and Xue, Y. Revisiting the Thrifty Gene Hypothesis via 65 Loci Associated with Susceptibility to Type 2 Diabetes. *The American Journal of Human Genetics*, 94(2) :176–185, Feb. 2014. ISSN 00029297. doi : 10.1016/j.ajhg.2013.12.010. URL http://linkinghub.elsevier.com/retrieve/pii/S0002929713005806.

Ballestar, E. Epigenetics Lessons from Twins : Prospects for Autoimmune Disease. *Clinical Reviews in Allergy & Immunology*, 39(1) :30–41, Aug. 2010. ISSN 1080-0549. doi : 10.1007/s12016-009-8168-4. URL http://dx.doi.org/10.1007/s12016-009-8168-4.

Bamshad, M. and Wooding, S. P. Signatures of natural selection in the human genome. *Nat Rev Genet*, 4(2) :99–111, Feb. 2003. ISSN 1471-0056. doi : 10.1038/nrg999. URL http://dx.doi.org/10.1038/nrg999.

Barreiro, L. B. and Quintana-Murci, L. From evolutionary genetics to human immunology : how selection shapes host defence genes. *Nat Rev Genet*, 11 (1) :17–30, Jan. 2010. ISSN 1471-0056. doi : 10.1038/nrg2698. URL http://dx.doi.org/10.1038/nrg2698.

Barreiro, L. B., Laval, G., Quach, H., Patin, E., and Quintana-Murci, L. Natural selection has driven population differentiation in modern humans. *Nat Genet*, 40 (3) :340–345, Mar. 2008. ISSN 1061-4036. doi : 10.1038/ng.78. URL http://dx.doi.org/10.1038/ng.78.

Barreiro, L. B., Ben-Ali, M., Quach, H., Laval, G., Patin, E., Pickrell, J. K., Bouchier, C., Tichit, M., Neyrolles, O., Gicquel, B., Kidd, J. R., Kidd, K. K., Alcaïs, A., Ragimbeau, J., Pellegrini, S., Abel, L., Casanova, J.-L., and Quintana-Murci, L. Evolutionary Dynamics of Human Toll-Like Receptors and Their Different Contributions to Host Defense. *PLoS Genet*, 5(7) :e1000562, 2009. doi : 10.1371/journal.pgen.1000562. URL http://dx.doi.org/10.1371%2Fjournal.pgen.1000562.

Barrow, T. M. and Michels, K. B. Epigenetic epidemiology of cancer. *Biochemical and Biophysical Research Communications*, 455(1-2) :70–83, Dec. 2014. ISSN 1090-2104. doi : 10.1016/j.bbrc.2014.08.002.

Batini, C., Lopes, J., Behar, D. M., Calafell, F., Jorde, L. B., van der Veen, L., Quintana-Murci, L., Spedini, G., Destro-Bisol, G., and Comas, D. Insights into the demographic history of African Pygmies from complete mitochondrial genomes. *Molecular Biology and Evolution*, 28(2) :1099–1110, Feb. 2011. ISSN 1537-1719. doi : 10.1093/molbev/msq294.

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and de Massy, B. PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. *Science*, 327(5967) :836–840, Feb. 2010. doi : 10.1126/science.1183439. URL http://www.sciencemag.org/content/327/5967/836.abstract.

Beall, C. M., Cavalleri, G. L., Deng, L., Elston, R. C., Gao, Y., Knight, J., Li, C., Li, J. C., Liang, Y., McCormack, M., Montgomery, H. E., Pan, H., Robbins, P. A., Shianna, K. V., Tam, S. C., Tsering, N., Veeramah, K. R., Wang, W., Wangdui, P., Weale, M. E., Xu, Y., Xu, Z., Yang, L., Zaman, M. J., Zeng, C., Zhang, L., Zhang, X., Zhaxi, P., and Zheng, Y. T. Natural selection on EPAS1 (HIF2) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National*

Academy of Sciences, 107(25) :11459–11464, June 2010. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.1002443107. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1002443107.

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., Gilad, Y., and Pritchard, J. K. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology*, 12(1) :R10, 2011. ISSN 1465-6914. doi : 10.1186/gb-2011-12-1-r10.

Bell, J. T., Tsai, P.-C., Yang, T.-P., Pidsley, R., Nisbet, J., Glass, D., Mangino, M., Zhai, G., Zhang, F., Valdes, A., Shin, S.-Y., Dempster, E. L., Murray, R. M., Grundberg, E., Hedman, A. K., Nica, A., Small, K. S., The MuTHER Consortium, Dermitzakis, E. T., McCarthy, M. I., Mill, J., Spector, T. D., and Deloukas, P. Epigenome-Wide Scans Identify Differentially Methylated Regions for Age and Age-Related Phenotypes in a Healthy Ageing Population. *PLoS Genetics*, 8(4) : e1002629, Apr. 2012. ISSN 1553-7404. doi : 10.1371/journal.pgen.1002629. URL http://dx.plos.org/10.1371/journal.pgen.1002629.

Bell, J. T., Loomis, A. K., Butcher, L. M., Gao, F., Zhang, B., Hyde, C. L., Sun, J., Wu, H., Ward, K., Harris, J., Scollen, S., Davies, M. N., Schalkwyk, L. C., Mill, J., MuTHER Consortium, Williams, F. M. K., Li, N., Deloukas, P., Beck, S., McMahon, S. B., Wang, J., John, S. L., and Spector, T. D. Differential methylation of the TRPA1 promoter in pain sensitivity. *Nature Communications*, 5 :2978, 2014. ISSN 2041-1723. doi : 10.1038/ncomms3978.

Berg, J. J. and Coop, G. A Population Genetic Signal of Polygenic Adaptation. *PLoS Genetics*, 10(8) :e1004412, Aug. 2014. ISSN 1553-7404. doi : 10.1371/journal.pgen.1004412. URL http://dx.plos.org/10.1371/journal.pgen.1004412.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S., and Thomson, J. A. The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, 28(10) :1045–1048, Oct. 2010. ISSN 1087-0156, 1546-1696. doi : 10.1038/nbt1010-1045. URL http://www.nature.com/doifinder/10.1038/nbt1010-1045.

Bernstein, E. and Allis, C. D. RNA meets chromatin. *Genes & Development*, 19(14) : 1635–1655, July 2005. ISSN 0890-9369. doi : 10.1101/gad.1324305.

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E., and Hirschhorn, J. N. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. *The American Journal of Human Genetics*, 74(6) :1111–1120, June 2004. ISSN 0002-9297. doi : 10.1086/421051. URL http://www.sciencedirect.com/science/article/pii/S0002929707628389.

Bhatia, G., Patterson, N., Pasaniuc, B., Zaitlen, N., Genovese, G., Pollack, S., Mallick, S., Myers, S., Tandon, A., Spencer, C., Palmer, C., Adeyemo, A., Akylbekova, E., Cupples, L., Divers, J., Fornage, M., Kao, W., Lange, L., Li, M., Musani, S., Mychaleckyj, J., Ogunniyi, A., Papanicolaou, G., Rotimi, C., Rotter, J., Ruczinski,

I., Salako, B., Siscovick, D., Tayo, B., Yang, Q., McCarroll, S., Sabeti, P., Lettre, G., DeJager, P., Hirschhorn, J., Zhu, X., Cooper, R., Reich, D., Wilson, J., and Price, A. Genome-wide Comparison of African-Ancestry Populations from CARe and Other Cohorts Reveals Signals of Natural Selection. *The American Journal of Human Genetics*, 89(3) :368–381, Sept. 2011. ISSN 00029297. doi : 10.1016/j.ajhg.2011.07.025. URL http://linkinghub.elsevier.com/retrieve/pii/S0002929711003545.

Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J. M., Mei, R., Scherer, S. W., Julian, C. G., Wilson, M. J., López Herráez, D., Brutsaert, T., Parra, E. J., Moore, L. G., and Shriver, M. D. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS genetics*, 6(9) :e1001116, Sept. 2010. ISSN 1553-7404. doi : 10.1371/journal.pgen.1001116.

Bigham, A. W., Mao, X., Mei, R., Brutsaert, T., Wilson, M. J., Julian, C. G., Parra, E. J., Akey, J. M., Moore, L. G., and Shriver, M. D. Identifying positive selection candidate loci for high-altitude adaptation in Andean populations. *Human Genomics*, 4(2) :79–90, Dec. 2009. ISSN 1479-7364.

Bjornsson, H. T., Sigurdsson, M. I., Fallin, M. D., Irizarry, R. A., Aspelund, T., Cui, H., Yu, W., Rongione, M. A., Ekström, T. J., Harris, T. B., Launer, L. J., Eiriksdottir, G., Leppert, M. F., Sapienza, C., Gudnason, V., and Feinberg, A. P. Intra-individual change over time in DNA methylation with familial clustering. *JAMA*, 299(24) : 2877–2883, June 2008. ISSN 1538-3598. doi : 10.1001/jama.299.24.2877.

Blekhman, R., Man, O., Herrmann, L., Boyko, A. R., Indap, A., Kosiol, C., Bustamante, C. D., Teshima, K. M., and Przeworski, M. Natural selection on genes that underlie human disease susceptibility. *Current biology : CB*, 18(12) :883–889, June 2008. ISSN 0960-9822. doi : 10.1016/j.cub.2008.04.074.

Boks, M. P., Derks, E. M., Weisenberger, D. J., Strengman, E., Janson, E., Sommer, I. E., Kahn, R. S., and Ophoff, R. A. The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PloS One*, 4(8) :e6767, 2009. ISSN 1932-6203. doi : 10.1371/journal.pone.0006767.

Boldsen, J. L. Leprosy in Medieval Denmark–osteological and epidemiological analyses. *Anthropologischer Anzeiger ; Bericht Über Die Biologisch-Anthropologische Literatur*, 67(4) :407–425, Dec. 2009. ISSN 0003-5548.

Boomsma, D., Busjahn, A., and Peltonen, L. Classical twin studies and beyond. *Nature Reviews Genetics*, 3(11) :872–882, Nov. 2002. ISSN 1471-0056, 1471-0064. doi : 10.1038/nrg932. URL http://www.nature.com/doifinder/10.1038/nrg932.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*, 132(2) :311–322, Jan. 2008. ISSN 00928674. doi : 10.1016/j.cell.2007.12.014. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867407016133.

Brown, C. J., Ballabio, A., Rupert, J. L., Lafreniere, R. G., Grompe, M., Tonlorenzi, R., and Willard, H. F. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*, 349(6304) :38–44, Jan. 1991. ISSN 0028-0836. doi : 10.1038/349038a0. URL http://www.nature.com/doifinder/10.1038/349038a0.

Bryc, K., Auton, A., Nelson, M. R., Oksenberg, J. R., Hauser, S. L., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C., Tishkoff, S. A., and Bustamante, C. D. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences of the United States of America*, 107(2) :786–791, Jan. 2010. ISSN 1091-6490. doi : 10.1073/pnas.0909559107.

Bryk, J., Hardouin, E., Pugach, I., Hughes, D., Strotmann, R., Stoneking, M., and Myles, S. Positive Selection in East Asians for an EDAR Allele that Enhances NF-B Activation. *PLoS ONE*, 3(5) :e2209, May 2008. ISSN 1932-6203. doi : 10.1371/journal.pone.0002209. URL http://dx.plos.org/10.1371/journal.pone.0002209.

Burger, L., Gaidatzis, D., Schübeler, D., and Stadler, M. B. Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Research*, 41(16) : e155, Sept. 2013. ISSN 1362-4962. doi : 10.1093/nar/gkt599.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D., Civello, D., Adams, M. D., Cargill, M., and Clark, A. G. Natural selection on protein-coding genes in the human genome. *Nature*, 437(7062) :1153–1157, Oct. 2005. ISSN 1476-4687. doi : 10.1038/nature04240.

Campbell, C. D., Chong, J. X., Malig, M., Ko, A., Dumont, B. L., Han, L., Vives, L., O'Roak, B. J., Sudmant, P. H., Shendure, J., Abney, M., Ober, C., and Eichler, E. E. Estimating the human mutation rate using autozygosity in a founder population. *Nat Genet*, 44(11) :1277–1281, Nov. 2012. ISSN 1061-4036. doi : 10.1038/ng.2418. URL http://dx.doi.org/10.1038/ng.2418.

Cann, R. L., Stoneking, M., and Wilson, A. C. Mitochondrial DNA and human evolution. *Nature*, 325(6099) :31–36, Jan. 1987. ISSN 0028-0836. doi : 10.1038/325031a0.

Cao, J. The functional role of long non-coding RNAs and epigenetics. *Biological Procedures Online*, 16(1) :11, 2014. ISSN 1480-9222. doi : 10.1186/1480-9222-16-11. URL http://www.biologicalproceduresonline.com/content/16/1/11.

Carlson, C. S., Thomas, D. J., Eberle, M. A., Swanson, J. E., Livingston, R. J., Rieder, M. J., and Nickerson, D. A. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Research*, 15(11) :1553–1565, Nov. 2005. ISSN 1088-9051. doi : 10.1101/gr.4326505.

Casals, F., Hodgkinson, A., Hussin, J., Idaghdour, Y., Bruat, V., de Maillard, T., Grenier, J.-C., Gbeha, E., Hamdan, F. F., Girard, S., Spinella, J.-F., Larivière, M., Saillour, V., Healy, J., Fernández, I., Sinnett, D., Michaud, J. L., Rouleau, G. A., Haddad, E., Le Deist, F., and Awadalla, P. Whole-Exome Sequencing Reveals a Rapid Change in the Frequency of Rare Functional Variants in a Founding Population of Humans. *PLoS Genetics*, 9(9) :e1003815, Sept. 2013. ISSN 1553-7404. doi : 10.1371/journal.pgen.1003815. URL http://dx.plos.org/10.1371/journal.pgen.1003815.

Casanova, J.-L., Abel, L., and Quintana-Murci, L. Immunology Taught by Human Genetics. *Cold Spring Harbor Symposia on Quantitative Biology*, 78(0) :157–172, Jan. 2013. ISSN 0091-7451, 1943-4456. doi : 10.1101/sqb.2013.78.019968. URL http://symposium.cshlp.org/cgi/doi/10.1101/sqb.2013.78.019968.

Castel, S. E. and Martienssen, R. A. RNA interference in the nucleus : roles for small RNAs in transcription, epigenetics and beyond. *Nature Reviews Genetics*, 14(2) : 100–112, Jan. 2013. ISSN 1471-0056, 1471-0064. doi : 10.1038/nrg3355. URL http://www.nature.com/doifinder/10.1038/nrg3355.

Cavalli-Sforza, L. L. Population structure and human evolution. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 164(995) :362–379, Mar. 1966. ISSN 0950-1193.

Cavalli-Sforza, L. L. and Bodmer, W. F. *The genetics of human populations.* WHFreeman, 1971.

Cavalli-Sforza, L. L. and Feldman, M. W. The application of molecular genetic approaches to the study of human evolution. *Nature Genetics*, 33 Suppl :266–275, Mar. 2003. ISSN 1061-4036. doi : 10.1038/ng1113.

Chakravarti, A. and Little, P. Nature, nurture and human disease. *Nature*, 421(6921) : 412–414, Jan. 2003. ISSN 0028-0836. doi : 10.1038/nature01401.

Charlesworth, B. The effects of deleterious mutations on evolution at linked sites. *Genetics*, 190(1) :5–22, Jan. 2012. ISSN 1943-2631. doi : 10.1534/genetics.111.134288.

Charlesworth, D. Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. *PLoS Genetics*, 2(4) :e64, 2006. ISSN 1553-7390, 1553-7404. doi : 10.1371/journal.pgen.0020064. URL http://dx.plos.org/10.1371/journal.pgen.0020064.

Chen, H., Patterson, N., and Reich, D. Population differentiation as a test for selective sweeps. *Genome Research*, 20(3) :393–402, Mar. 2010. ISSN 1549-5469. doi : 10.1101/gr.100545.109.

Chen, H., Hey, J., and Slatkin, M. A hidden Markov model for investigating recent positive selection through haplotype structure. *Theoretical Population Biology*, 99 : 18–30, Feb. 2015. ISSN 00405809. doi : 10.1016/j.tpb.2014.11.001. URL http://linkinghub.elsevier.com/retrieve/pii/S0040580914000914.

Chen, Y. S., Torroni, A., Excoffier, L., Santachiara-Benerecetti, A. S., and Wallace, D. C. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *American Journal of Human Genetics*, 57(1) :133–149, July 1995. ISSN 0002-9297.

Chevin, L.-M. and Hospital, F. Selective sweep at a quantitative trait locus in the presence of background genetic variation. *Genetics*, 180(3) :1645–1660, Nov. 2008. ISSN 0016-6731. doi : 10.1534/genetics.108.093351.

Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055) : 69–87, Sept. 2005. ISSN 1476-4687. doi : 10.1038/nature04072.

Christensen, B. C., Houseman, E. A., Marsit, C. J., Zheng, S., Wrensch, M. R., Wiemels, J. L., Nelson, H. H., Karagas, M. R., Padbury, J. F., Bueno, R., Sugarbaker, D. J., Yeh, R.-F., Wiencke, J. K., and Kelsey, K. T. Aging and Environmental Exposures Alter Tissue-Specific DNA Methylation Dependent upon CpG Island Context. *PLoS Genetics*, 5(8) :e1000602, Aug. 2009. ISSN 1553-7404. doi : 10.1371/journal.pgen.1000602. URL http://dx.plos.org/10.1371/journal.pgen.1000602.

Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M. F., Kellis, M., Lindblad-Toh, K., and Lander, E. S. Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences*, 104(49) : 19428–19433, Dec. 2007. doi : 10.1073/pnas.0709013104. URL http://www.pnas.org/content/104/49/19428.abstract.

Clark, A. G. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*, 15(11) :1496–1502, Nov. 2005. ISSN 1088-9051. doi : 10.1101/gr.4107905. URL http://www.genome.org/cgi/doi/10.1101/gr.4107905.

Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P., Kejariwal, A., Todd, M. J., Tanenbaum, D. M., Civello, D., Lu, F., Murphy, B., Ferriera, S., Wang, G., Zheng, X., White, T. J., Sninsky, J. J., Adams, M. D., and Cargill, M. Positive selection in the human genome inferred from human-chimp-mouse orthologous gene alignments. *Cold Spring Harbor Symposia on Quantitative Biology*, 68 :471–477, 2003. ISSN 0091-7451.

Colot, V. and Rossignol, J. L. Eukaryotic DNA methylation as an evolutionary device. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, 21(5) :402–411, May 1999. ISSN 0265-9247. doi : 10.1002/(SICI)1521-1878(199905)21:5<402::AID-BIES7>3.0.CO;2-B.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., MacArthur, D. G., MacDonald, J. R., Onyiah, I., Pang, A. W. C., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Tyler-Smith, C., Carter, N. P., Lee, C., Scherer, S. W., and Hurles, M. E. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289) :704–712, Apr.

2010. ISSN 0028-0836. doi : 10.1038/nature08516. URL http://dx.doi.org/10.1038/nature08516.

Consortium, I. H. G. S. Initial sequencing and analysis of the human genome. *Nature*, 409(6822) :860–921, Feb. 2001. ISSN 0028-0836. doi : 10.1038/35057062. URL http://dx.doi.org/10.1038/35057062.

Coop, G., Pickrell, J. K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R. M., Cavalli-Sforza, L. L., Feldman, M. W., and Pritchard, J. K. The role of geography in human adaptation. *PLoS genetics*, 5(6) :e1000500, June 2009. ISSN 1553-7404. doi : 10.1371/journal.pgen.1000500.

Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J. K. Using Environmental Correlations to Identify Loci Underlying Local Adaptation. *Genetics*, 185(4) : 1411–1423, Aug. 2010. ISSN 0016-6731. doi : 10.1534/genetics.110.114819. URL http://www.genetics.org/cgi/doi/10.1534/genetics.110.114819.

Cooper, G. M., Nickerson, D. A., and Eichler, E. E. Mutational and selective effects on copy-number variants in the human genome. *Nat Genet*, June 2007.

Crawford, J. E. and Lazzaro, B. P. Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. *Frontiers in Genetics*, 3 :66, 2012. ISSN 1664-8021. doi : 10.3389/fgene.2012.00066.

Crispo, E. THE BALDWIN EFFECT AND GENETIC ASSIMILATION : REVISITING TWO MECHANISMS OF EVOLUTIONARY CHANGE MEDIATED BY PHENOTYPIC PLASTICITY. *Evolution*, 61(11) :2469–2479, Nov. 2007. ISSN 0014-3820, 1558-5646. doi : 10.1111/j.1558-5646.2007.00203.x. URL http://doi.wiley.com/10.1111/j.1558-5646.2007.00203.x.

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Kähäri, A. K., Keenan, S., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Aken, B. L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S. M., Spudich, G., Trevanion, S. J., Yates, A., Zerbino, D. R., and Flicek, P. Ensembl 2015. *Nucleic Acids Research*, Oct. 2014. doi : 10.1093/nar/gku1010. URL http://nar.oxfordjournals.org/content/early/2014/10/28/nar.gku1010.abstract.

Cutter, A. D. and Payseur, B. A. Genomic signatures of selection at linked sites : unifying the disparity among species. *Nature Reviews Genetics*, 14(4) :262–274, Mar. 2013. ISSN 1471-0056, 1471-0064. doi : 10.1038/nrg3425. URL http://www.nature.com/doifinder/10.1038/nrg3425.

Dang, M. N., Buzzetti, R., and Pozzilli, P. Epigenetics in autoimmune diseases with focus on type 1 diabetes. *Diabetes/Metabolism Research and Reviews*, 29(1) :8–18, Jan. 2013. ISSN 1520-7560. doi : 10.1002/dmrr.2375.

Darwin, C. *On the origin of species*. John Murray, 1859.

Daub, J. T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M., and Excoffier, L. Evidence for polygenic adaptation to pathogens in the human genome. *Molecular Biology and Evolution*, 30(7) :1544–1558, July 2013. ISSN 1537-1719. doi : 10.1093/molbev/mst080.

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology*, 6(12) :e1001025, 2010. ISSN 1553-7358. doi : 10.1371/journal.pcbi.1001025.

Dick, D. M. and Foroud, T. Candidate Genes for Alcohol Dependence : A Review of Genetic Evidence From Human Studies :. *Alcoholism : Clinical & Experimental Research*, 27(5) :868–879, May 2003. ISSN 0145-6008. doi : 10.1097/01.ALC.0000065436.24221.63. URL http://doi.wiley.com/10.1097/01.ALC.0000065436.24221.63.

Diller, K. C., Gilbert, W. A., and Kocher, T. D. Selective Sweeps in the Human Genome : A Starting Point for Identifying Genetic Differences Between Modern Humans and Chimpanzees. *Molecular Biology and Evolution*, 19(12) :2342–2345, Dec. 2002. URL http://mbe.oxfordjournals.org/content/19/12/2342.short.

Ding, Q., Hu, Y., Xu, S., Wang, J., and Jin, L. Neanderthal Introgression at Chromosome 3p21.31 Was Under Positive Natural Selection in East Asians. *Molecular Biology and Evolution*, 31(3) :683–695, Mar. 2014. ISSN 0737-4038, 1537-1719. doi : 10.1093/molbev/mst260. URL http://mbe.oxfordjournals.org/cgi/doi/10.1093/molbev/mst260.

Do, R., Balick, D., Li, H., Adzhubei, I., Sunyaev, S., and Reich, D. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat Genet*, 47(2) :126–131, Feb. 2015. ISSN 1061-4036. URL http://dx.doi.org/10.1038/ng.3186.

Dolinoy, D., Weidman, J., and Jirtle, R. Epigenetic gene regulation : Linking early developmental environment to adult disease. *Reproductive Toxicology*, 23(3) :297–307, Apr. 2007. ISSN 08906238. doi : 10.1016/j.reprotox.2006.08.012. URL http://linkinghub.elsevier.com/retrieve/pii/S0890623806001973.

Drake, A. J., O'Shaughnessy, P. J., Bhattacharya, S., Monteiro, A., Kerrigan, D., Goetz, S., Raab, A., Rhind, S. M., Sinclair, K. D., Meharg, A. A., Feldmann, J., and Fowler, P. A. In utero exposure to cigarette chemicals induces sex-specific disruption of one-carbon metabolism and DNA methylation in the human fetal liver. *BMC Medicine*, 13(1), Dec. 2015. ISSN 1741-7015. doi : 10.1186/s12916-014-0251-x. URL http://www.biomedcentral.com/1741-7015/13/18.

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B., Yeung, G., Dahl, F., Fernandez,

A., Staker, B., Pant, K. P., Baccash, J., Borcherding, A. P., Brownley, A., Cedeno, R., Chen, L., Chernikoff, D., Cheung, A., Chirita, R., Curson, B., Ebert, J. C., Hacker, C. R., Hartlage, R., Hauser, B., Huang, S., Jiang, Y., Karpinchyk, V., Koenig, M., Kong, C., Landers, T., Le, C., Liu, J., McBride, C. E., Morenzoni, M., Morey, R. E., Mutch, K., Perazich, H., Perry, K., Peters, B. A., Peterson, J., Pethiyagoda, C. L., Pothuraju, K., Richter, C., Rosenbaum, A. M., Roy, S., Shafto, J., Sharanhovich, U., Shannon, K. W., Sheppy, C. G., Sun, M., Thakuria, J. V., Tran, A., Vu, D., Zaranek, A. W., Wu, X., Drmanac, S., Oliphant, A. R., Banyai, W. C., Martin, B., Ballinger, D. G., Church, G. M., and Reid, C. A. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science (New York, N.Y.)*, 327 (5961) :78–81, Jan. 2010. ISSN 1095-9203. doi : 10.1126/science.1181498.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B.-K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Dunham, I., Ernst, J., Furey, T. S., Gerstein, M., Giardine, B., Greven, M., Hardison, R. C., Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kellis, M., Khatun, J., Kheradpour, P., Kundaje, A., Lassmann, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., Mortazavi, A., Parker, S. C. J., Reddy, T. E., Rozowsky, J., Schlesinger, F., Thurman, R. E., Wang, J., Ward, L. D., Whitfield, T. W., Wilder, S. P., Wu, W., Xi, H. S., Yip, K. Y., Zhuang, J., Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., Snyder, M., Pazin, M. J., Lowdon, R. F., Dillon, L. A. L., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Green, E. D., Good, P. J., Feingold, E. A., Bernstein, B. E., Birney, E., Crawford, G. E., Dekker, J., Elnitski, L., Farnham, P. J., Gerstein, M., Giddings, M. C., Gingeras, T. R., Green, E. D., Guigó, R., Hardison, R. C., Hubbard, T. J., Kellis, M., Kent, W. J., Lieb, J. D., Margulies, E. H., Myers, R. M., Snyder, M., Stamatoyannopoulos, J. A., Tenenbaum, S. A., Weng, Z., White, K. P., Wold, B., Khatun, J., Yu, Y., Wrobel, J., Risk, B. A., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Chen, X., Giddings, M. C., Bernstein, B. E., Epstein, C. B., Shoresh, N., Ernst, J., Kheradpour, P., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Ward, L. D., Altshuler, R. C., Eaton, M. L., Kellis, M., Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H. P., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O. J., Park, E., Preall, J. B., Presaud, K., Ribeca, P., Risk, B. A., Robyr, D., Ruan, X., Sammeth, M., Sandhu, K. S., Schaeffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T. J., Reymond, A., Antonarakis, S. E., Hannon, G. J., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P.,

Guigó, R., Gingeras, T. R., Rosenbloom, K. R., Sloan, C. A., Learned, K., Malladi, V. S., Wong, M. C., Barber, G. P., Cline, M. S., Dreszer, T. R., Heitner, S. G., Karolchik, D., Kent, W. J., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Furey, T. S., Song, L., Grasfeder, L. L., Giresi, P. G., Lee, B.-K., Battenhouse, A., Sheffield, N. C., Simon, J. M., Showers, K. A., Safi, A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Ki Kim, S., Zhang, Z. Z., Mieczkowski, P. A., Mieczkowska, J. O., Liu, Z., McDaniell, R. M., Ni, Y., Rashid, N. U., Kim, M. J., Adar, S., Zhang, Z., Wang, T., Winter, D., Keefe, D., Birney, E., Iyer, V. R., Lieb, J. D., Crawford, G. E., Li, G., Sandhu, K. S., Zheng, M., Wang, P., Luo, O. J., Shahab, A., Fullwood, M. J., Ruan, X., Ruan, Y., Myers, R. M., Pauli, F., Williams, B. A., Gertz, J., Marinov, G. K., Reddy, T. E., Vielmetter, J., Partridge, E., Trout, D., Varley, K. E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K. M., Anaya, M., Cross, M. K., King, B., Muratet, M. A., Antoshechkin, I., Newberry, K. M., McCue, K., Nesmith, A. S., Fisher-Aylor, K. I., Pusey, B., DeSalvo, G., Parker, S. L., Balasubramanian, S., Davis, N. S., Meadows, S. K., Eggleston, T., Gunter, C., Newberry, J. S., Levy, S. E., Absher, D. M., Mortazavi, A., Wong, W. H., Wold, B., Blow, M. J., Visel, A., Pennachio, L. A., Elnitski, L., Margulies, E. H., Parker, S. C. J., Petrykowska, H. M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Chrast, J., Davidson, C., Derrien, T., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J. M., Griffiths, E., Harte, R., Hendrix, D. A., Howald, C., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Kokocinski, F., Leng, J., Lin, M. F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J. M., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward, C., Tanzer, A., Tapanari, E., Tress, M. L., van Baren, M. J., Walters, N., Washie. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414) :57–74, Sept. 2012. ISSN 0028-0836, 1476-4687. doi : 10.1038/nature11247. URL http://www.nature.com/doifinder/10.1038/nature11247.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., and Beck, S. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, 38 (12) :1378–1385, Dec. 2006. ISSN 1061-4036. doi : 10.1038/ng1909. URL http://www.nature.com/doifinder/10.1038/ng1909.

Edmonds, C. A., Lillie, A. S., and Cavalli-Sforza, L. L. Mutations arising in the wave front of an expanding population. *Proceedings of the National Academy of Sciences of the United States of America*, 101(4) :975–979, Jan. 2004. ISSN 0027-8424. doi : 10.1073/pnas.0308064100.

Ehrlich, M., Gama-Sosa, M. A., Huang, L.-H., Midgett, R. M., Kuo, K. C., McCune, R. A., and Gehrke, C. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic Acids Research*, 10(8) :2709–2721, 1982. ISSN 0305-1048, 1362-4962. doi : 10.1093/nar/10.8.2709. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/10.8.2709.

Eilertson, K. E., Booth, J. G., and Bustamante, C. D. SnIPRE : Selection Inference Using a Poisson Random Effects Model. *PLoS Computational Biology*, 8(12) : e1002806, Dec. 2012. ISSN 1553-7358. doi : 10.1371/journal.pcbi.1002806. URL http://dx.plos.org/10.1371/journal.pcbi.1002806.

Enard, D., Messer, P. W., and Petrov, D. A. Genome-wide signals of positive selection in human evolution. *Genome Research*, 24(6) :885–895, June 2014. doi : 10.1101/gr.164822.113. URL http://genome.cshlp.org/content/24/6/885.abstract.

Enattah, N. S., Sahi, T., Savilahti, E., Terwilliger, J. D., Peltonen, L., and Järvelä, I. Identification of a variant associated with adult-type hypolactasia. *Nature Genetics*, 30(2) :233–237, Feb. 2002. ISSN 10614036. doi : 10.1038/ng826. URL http://www.nature.com/doifinder/10.1038/ng826.

ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)*, 306(5696) :636–640, Oct. 2004. ISSN 1095-9203. doi : 10.1126/science.1105136.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414) :57–74, Sept. 2012. ISSN 1476-4687. doi : 10.1038/nature11247.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345) :43–49, May 2011. ISSN 0028-0836. doi : 10.1038/nature09906. URL http://dx.doi.org/10.1038/nature09906.

Essex, M. J., Thomas Boyce, W., Hertzman, C., Lam, L. L., Armstrong, J. M., Neumann, S. M. A., and Kobor, M. S. Epigenetic Vestiges of Early Developmental Adversity : Childhood Stress Exposure and DNA Methylation in Adolescence : **Epigenetic Vestiges of Early Adversity**. *Child Development*, 84(1) :58–75, Jan. 2013. ISSN 00093920. doi : 10.1111/j.1467-8624.2011.01641.x. URL http://doi.wiley.com/10.1111/j.1467-8624.2011.01641.x.

Excoffier, L. Human demographic history : refining the recent African origin model. *Current Opinion in Genetics & Development*, 12(6) :675–682, Dec. 2002. ISSN 0959437X. doi : 10.1016/S0959-437X(02)00350-7. URL http://linkinghub.elsevier.com/retrieve/pii/S0959437X02003507.

Excoffier, L., Smouse, P. E., and Quattro, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes : application to human mitochondrial DNA restriction data. *Genetics*, 131(2) :479–491, June 1992. ISSN 0016-6731.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., and Foll, M. Robust demographic inference from genomic and SNP data. *PLoS genetics*, 9(10) : e1003905, Oct. 2013. ISSN 1553-7404. doi : 10.1371/journal.pgen.1003905.

Eyre-Walker, A. and Keightley, P. D. Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change. *Molecular Biology and Evolution*, 26(9) :2097–2108, Sept. 2009. doi : 10.1093/molbev/msp119. URL http://mbe.oxfordjournals.org/content/26/9/2097.abstract.

Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., and Tress, M. L. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Human Molecular Genetics*, 23(22) :5866–5878, Nov. 2014. ISSN 1460-2083. doi : 10.1093/hmg/ddu309.

Fagundes, N. J. R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F. M., Bonatto, S. L., and Excoffier, L. Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences*, 104(45) : 17614 –17619, Nov. 2007. doi : 10.1073/pnas.0708280104. URL http://www.pnas.org/content/104/45/17614.abstract.

Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., and Servin, B. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics*, 193(3) :929–941, Mar. 2013. ISSN 1943-2631. doi : 10.1534/genetics.112.147231.

Fay, J. C. and Wu, C. I. Hitchhiking under positive Darwinian selection. *Genetics*, 155(3) :1405–1413, July 2000. ISSN 0016-6731.

Feinberg, A. P. Phenotypic plasticity and the epigenetics of human disease. *Nature*, 447(7143) :433–440, May 2007. ISSN 0028-0836, 1476-4687. doi : 10.1038/nature05919. URL http://www.nature.com/doifinder/10.1038/nature05919.

Feinberg, A. P. and Irizarry, R. A. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proceedings of the National Academy of Sciences*, 107(suppl_1) :1757–1764, Jan. 2010. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.0906183107. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0906183107.

Felsenfeld, G. and Groudine, M. Controlling the double helix. *Nature*, 421(6921) : 448–453, Jan. 2003. ISSN 0028-0836. doi : 10.1038/nature01411.

Feng, S., Jacobsen, S. E., and Reik, W. Epigenetic Reprogramming in Plant and Animal Development. *Science*, 330(6004) :622–627, Oct. 2010. ISSN 0036-8075, 1095-9203. doi : 10.1126/science.1190614. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1190614.

Feuk, L., Carson, A. R., and Scherer, S. W. Structural variation in the human genome. *Nat Rev Genet*, 7(2) :85–97, Feb. 2006. ISSN 1471-0056. doi : 10.1038/nrg1767. URL http://dx.doi.org/10.1038/nrg1767.

Foll, M., Gaggiotti, O. E., Daub, J. T., Vatsiou, A., and Excoffier, L. Widespread signals of convergent adaptation to high altitude in Asia and america. *American Journal of Human Genetics*, 95(4) :394–407, Oct. 2014. ISSN 1537-6605. doi : 10.1016/j.ajhg.2014.09.002.

Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., Heine-Suner, D., Cigudosa, J. C., Urioste, M., Benitez, J., Boix-Chornet, M., Sanchez-Aguilera, A., Ling, C., Carlsson, E., Poulsen, P., Vaag, A., Stephan, Z., Spector, T. D., Wu, Y.-Z., Plass, C., and Esteller, M. From The Cover : Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences*, 102(30) :10604–10609, July 2005. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.0500398102. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0500398102.

Fraser, H., Lam, L., Neumann, S., and Kobor, M. Population-specificity of human DNA methylation. *Genome Biology*, 13(2) :R8, 2012. URL http://genomebiology.com/2012/13/2/R8.

Fraser, H. B. Gene expression drives local adaptation in humans. *Genome Research*, 23(7) :1089–1096, July 2013. ISSN 1088-9051. doi : 10.1101/gr.152710.112. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.152710.112.

Friedberg, E. C. DNA damage and repair. *Nature*, 421(6921) :436–440, Jan. 2003. ISSN 0028-0836. doi : 10.1038/nature01408.

Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S. M., Bondarev, A. A., Johnson, P. L. F., Aximu-Petri, A., Prufer, K., de Filippo, C., Meyer, M., Zwyns, N., Salazar-Garcia, D. C., Kuzmin, Y. V., Keates, S. G., Kosintsev, P. A., Razhev, D. I., Richards, M. P., Peristov, N. V., Lachmann, M., Douka, K., Higham, T. F. G., Slatkin, M., Hublin, J.-J., Reich, D., Kelso, J., Viola, T. B., and Paabo, S. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, 514(7523) :445–449, Oct. 2014. ISSN 0028-0836. URL http://dx.doi.org/10.1038/nature13810.

Fu, W. and Akey, J. M. Selection and Adaptation in the Human Genome. *Annual Review of Genomics and Human Genetics*, 14 (1) :467–489, Aug. 2013. ISSN 1527-8204, 1545-293X. doi : 10.1146/annurev-genom-091212-153509. URL http://www.annualreviews.org/doi/abs/10.1146/annurev-genom-091212-153509.

Fu, Y. X. and Li, W. H. Statistical Tests of Neutrality of Mutations. *Genetics*, 133(3) : 693 –709, Mar. 1993. URL http://www.genetics.org/content/133/3/693.abstract.

Fujimoto, A., Kimura, R., Ohashi, J., Omi, K., Yuliwulandari, R., Batubara, L., Mustofa, M. S., Samakkarn, U., Settheetham-Ishida, W., Ishida, T., Morishita, Y., Furusawa, T., Nakazawa, M., Ohtsuka, R., and Tokunaga, K. A scan for genetic determinants of human hair morphology : EDAR is associated with Asian hair thickness. *Human Molecular Genetics*, 17(6) :835–843, Mar. 2008. ISSN 1460-2083. doi : 10.1093/hmg/ddm355.

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admetlla, A., Ferrer-Admettla, A., Pattini, L., and Nielsen, R. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS genetics*, 7 (11) :e1002355, Nov. 2011. ISSN 1553-7404. doi : 10.1371/journal.pgen.1002355.

Galinsky, K. J., Bhatia, G., Loh, P.-R., Georgiev, S., Mukherjee, S., Patterson, N. J., and Price, A. L. Fast principal components analysis reveals independent evolution of ADH1b gene in Europe and East Asia. Technical Report biorxiv ;018143v2, Apr. 2015. URL http://biorxiv.org/lookup/doi/10.1101/018143.

Gertz, J., Varley, K. E., Reddy, T. E., Bowling, K. M., Pauli, F., Parker, S. L., Kucera, K. S., Willard, H. F., and Myers, R. M. Analysis of DNA Methylation in a Three-Generation Family Reveals Widespread Genetic Influence on Epigenetic Regulation. *PLoS Genetics*, 7(8) :e1002228, Aug. 2011. ISSN 1553-7404. doi : 10.1371/journal.pgen.1002228. URL http://dx.plos.org/10.1371/journal.pgen.1002228.

Gibbs, J. R., van der Brug, M. P., Hernandez, D. G., Traynor, B. J., Nalls, M. A., Lai, S.-L., Arepalli, S., Dillman, A., Rafferty, I. P., Troncoso, J., Johnson, R., Zielke, H. R., Ferrucci, L., Longo, D. L., Cookson, M. R., and Singleton, A. B. Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain. *PLoS Genetics*, 6(5) :e1000952, May 2010. ISSN 1553-7404. doi : 10.1371/journal.pgen.1000952. URL http://dx.plos.org/10.1371/journal.pgen.1000952.

Gilad, Y. and Lancet, D. Population differences in the human functional olfactory repertoire. *Molecular Biology and Evolution*, 20(3) :307–314, Mar. 2003. ISSN 0737-4038.

Gilad, Y., Bustamante, C. D., Lancet, D., and Pääbo, S. Natural selection on the olfactory receptor gene family in humans and chimpanzees. *American Journal of Human Genetics*, 73(3) :489–501, Sept. 2003. ISSN 0002-9297. doi : 10.1086/378132.

Gokhman, D., Lavi, E., Prufer, K., Fraga, M. F., Riancho, J. A., Kelso, J., Paabo, S., Meshorer, E., and Carmel, L. Reconstructing the DNA Methylation Maps of the Neandertal and the Denisovan. *Science*, 344(6183) :523–527, May 2014. ISSN 0036-8075, 1095-9203. doi : 10.1126/science.1250368. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1250368.

Goldberg, A. D., Allis, C. D., and Bernstein, E. Epigenetics : A Landscape Takes Shape. *Cell*, 128(4) :635–638, Feb. 2007. ISSN 00928674. doi : 10.1016/j.cell.2007.02.006. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867407001869.

Granka, J. M., Henn, B. M., Gignoux, C. R., Kidd, J. M., Bustamante, C. D., and Feldman, M. W. Limited Evidence for Classic Selective Sweeps in African Populations. *Genetics*, 192(3) :1049–1064, Nov. 2012. ISSN 0016-6731. doi : 10.1534/genetics.112.144071. URL http://www.genetics.org/cgi/doi/10.1534/genetics.112.144071.

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A., The 1000 Genomes Project, Bustamante, C. D., Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., Nickerson, D. A., Peltonen, L., Schafer, A. J., Sherry, S. T., Wang, J., Wilson, R. K., Gibbs, R. A., Deiros, D., Metzker, M., Muzny, D., Reid, J., Wheeler, D., Wang, J., Li, J., Jian, M., Li, G., Li, R., Liang, H., Tian, G., Wang, B., Wang, J., Wang, W., Yang, H., Zhang, X., Zheng, H., Lander, E. S., Altshuler, D. L., Ambrogio, L., Bloom, T., Cibulskis, K., Fennell, T. J., Gabriel, S. B., Jaffe, D. B., Shefler, E., Sougnez, C. L., Bentley, D. R., Gormley, N., Humphray, S., Kingsbury, Z., Koko-Gonzales, P., Stone, J., McKernan, K. J., Costa, G. L., Ichikawa, J. K., Lee, C. C., Sudbrak, R., Lehrach, H., Borodina, T. A., Dahl, A., Davydov, A. N., Marquardt, P., Mertes, F., Nietfeld, W., Rosenstiel, P., Schreiber, S., Soldatov, A. V., Timmermann, B., Tolzmann, M., Egholm, M., Affourtit, J., Ashworth, D., Attiya, S., Bachorski, M., Buglione, E., Burke, A., Caprio, A., Celone, C., Clark, S., Conners, D., Desany, B., Gu, L., Guccione, L., Kao, K., Kebbel, A., Knowlton, J., Labrecque, M., McDade, L., Mealmaker, C., Minderman, M., Nawrocki, A., Niazi, F., Pareja, K., Ramenani, R., Riches, D., Song, W., Turcotte, C., Wang, S., Mardis, E. R., Wilson, R. K., Dooling, D., Fulton, L., Fulton, R., Weinstock, G., Durbin, R. M., Burton, J., Carter, D. M., Churcher, C., Coffey, A., Cox, A., Palotie, A., Quail, M., Skelly, T., Stalker, J., Swerdlow, H. P., Turner, D., De Witte, A., Giles, S., Gibbs, R. A., Wheeler, D., Bainbridge, M., Challis, D., Sabo, A., Yu, F., Yu, J., Wang, J., Fang, X., Guo, X., Li, R., Li, Y., Luo, R., Tai, S., Wu, H., Zheng, H., Zheng, X., Zhou, Y., Li, G., Wang, J., Yang, H., Marth, G. T., Garrison, E. P., Huang, W., Indap, A., Kural, D., Lee, W.-P., Leong, W. F., Quinlan, A. R., Stewart, C., Stromberg, M. P., Ward, A. N., Wu, J., Lee, C., Mills, R. E., Shi, X., Daly, M. J., DePristo, M. A., Altshuler, D. L., Ball, A. D., Banks, E., Bloom, T., Browning, B. L., Cibulskis, K., Fennell, T. J., Garimella, K. V., Grossman, S. R., Handsaker, R. E., Hanna, M., Hartl, C., Jaffe, D. B., Kernytsky, A. M., Korn, J. M., Li, H., Maguire, J. R., McCarroll, S. A., McKenna, A., Nemesh, J. C., Philippakis, A. A., Poplin, R. E., Price, A., Rivas, M. A., Sabeti, P. C., Schaffner, S. F., Shefler, E., Shlyakhter, I. A., Cooper, D. N., Ball, E. V., Mort, M., Phillips, A. D., Stenson, P. D., Sebat, J., Makarov, V., Ye, K., Yoon, S. C., Bustamante, C. D., Clark, A. G., Boyko, A., Degenhardt, J., Gravel, S., Gutenkunst, R. N., Kaganovich, M., Keinan, A., Lacroute, P., Ma, X., Reynolds, A., Clarke, L., Flicek, P., Cunningham, F., Herrero, J., Keenen, S., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Smith, R. E., Zalunin, V., Zheng-Bradley, X., Korbel, J. O., Stutz, A. M., Humphray, S., Bauer, M., Cheetham, R. K., Cox, T., Eberle, M., James, T., Kahn, S., Murray, L., Chakravarti, A., Ye, K., De La Vega, F. M., Fu, Y., Hyland, F. C. L., Manning, J. M., McLaughlin, S. F., Peckham, H. E., Sakarya, O., Sun, Y. A., Tsung, E. F., Batzer, M. A., Konkel, M. K., Walker, J. A., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Herwig, R., Parkhomchuk, D. V., Sherry, S. T., Agarwala, R., Khouri, H. M., Morgulis, A. O., Paschall, J. E., Phan, L. D., Rotmistrovsky, K. E., Sanders, R. D., Shumway, M. F., Xiao, C., McVean, G. A., Auton, A., Iqbal, Z., Lunter, G., Marchini, J. L., Moutsianas, L., Myers, S., Tumian, A., Desany, B., Knight, J., Winer, R., Craig, D. W., Beckstrom-Sternberg, S. M., Christoforides, A., Kurdoglu,

A. A., Pearson, J. V., Sinari, S. A., Tembe, W. D., Haussler, D., Hinrichs, A. S., Katzman, S. J., Kern, A., Kuhn, R. M., Przeworski, M., Hernandez, R. D., Howie, B., Kelley, J. L., Melton, S. C., Abecasis, G. R., Li, Y., Anderson, P., Blackwell, T., Chen, W., Cookson, W. O., Ding, J., Kang, H. M., Lathrop, M., Liang, L., Moffatt, M. F., Scheet, P., Sidore, C., Snyder, M., Zhan, X., Zollner, S., Awadalla, P., Casals, F., Idaghdour, Y., Keebler, J., Stone, E. A., Zilversmit, M., Jorde, L., Xing, J., Eichler, E. E., Aksay, G., Alkan, C., Hajirasouliha, I., Hormozdiari, F., Kidd, J. M., Sahinalp, S. C., Sudmant, P. H., Mardis, E. R., Chen, K., Chinwalla, A., Ding, L., Koboldt, D. C., McLellan, M. D., Dooling, D., Weinstock, G., Wallis, J. W., Wendl, M. C., Zhang, Q., Durbin, R. M., Albers, C. A., Ayub, Q., Balasubramaniam, S., Barrett, J. C., Carter, D. M., Chen, Y., Conrad, D. F., Danecek, P., Dermitzakis, E. T., Hu, M., Huang, N., Hurles, M. E., Jin, H., Jostins, L., Keane, T. M., Le, S. Q., Lindsay, S., Long, Q., MacArthur, D. G., Montgomery, S. B., Parts, L., Stalker, J., Tyler-Sm. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29) :11983–11988, July 2011. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.1019276108. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1019276108.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L. F., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., and Pääbo, S. A draft sequence of the Neandertal genome. *Science (New York, N.Y.)*, 328(5979) :710–722, May 2010. ISSN 1095-9203. doi : 10.1126/science.1188021.

Grossman, S., Andersen, K., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D., Griesemer, D., Karlsson, E., Wong, S., Cabili, M., Adegbola, R., Bamezai, R., Hill, A., Vannberg, F., Rinn, J., Lander, E., Schaffner, S., and Sabeti, P. Identifying Recent Adaptations in Large-Scale Genomic Data. *Cell*, 152(4) : 703–713, Feb. 2013. ISSN 00928674. doi : 10.1016/j.cell.2013.01.035. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867413000871.

Grossman, S. R., Shlyakhter, I., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., Lander, E. S., Schaffner, S. F., and Sabeti, P. C. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science (New York, N.Y.)*, 327 (5967) :883–886, Feb. 2010. ISSN 1095-9203. doi : 10.1126/science.1183863.

Grundberg, E., Small, K. S., Hedman, K., Nica, A. C., Buil, A., Keildson, S., Bell, J. T., Yang, T.-P., Meduri, E., Barrett, A., Nisbett, J., Sekowska, M., Wilk, A., Shin, S.-Y., Glass, D., Travers, M., Min, J. L., Ring, S., Ho, K., Thorleifsson, G., Kong, A., Thorsteindottir, U., Ainali, C., Dimas, A. S., Hassanali, N., Ingle, C., Knowles, D., Krestyaninova, M., Lowe, C. E., Di Meglio, P., Montgomery,

S. B., Parts, L., Potter, S., Surdulescu, G., Tsaprouni, L., Tsoka, S., Bataille, V., Durbin, R., Nestle, F. O., O'Rahilly, S., Soranzo, N., Lindgren, C. M., Zondervan, K. T., Ahmadi, K. R., Schadt, E. E., Stefansson, K., Smith, G. D., McCarthy, M. I., Deloukas, P., Dermitzakis, E. T., and Spector, T. D. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics*, 44(10) :1084–1089, Sept. 2012. ISSN 1061-4036, 1546-1718. doi : 10.1038/ng.2394. URL http://www.nature.com/doifinder/10.1038/ng.2394.

Gu, W., Zhang, F., and Lupski, J. R. Mechanisms for human genomic rearrangements. *PathoGenetics*, 1(1) :4, 2008. ISSN 1755-8417. doi : 10.1186/1755-8417-1-4.

Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S. B., Buil, A., Ongen, H., Yurovsky, A., Bryois, J., Giger, T., Romano, L., Planchon, A., Falconnet, E., Bielser, D., Gagnebin, M., Padioleau, I., Borel, C., Letourneau, A., Makrythanasis, P., Guipponi, M., Gehrig, C., Antonarakis, S. E., and Dermitzakis, E. T. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife*, 2, June 2013. ISSN 2050-084X. doi : 10.7554/eLife.00523. URL http://elifesciences.org/lookup/doi/10.7554/eLife.00523.

Ha, H., Song, J., Wang, S., Kapusta, A., Feschotte, C., Chen, K. C., and Xing, J. A comprehensive analysis of piRNAs from adult human testis and their relationship with genes and mobile elements. *BMC Genomics*, 15(1) :545, 2014. ISSN 1471-2164. doi : 10.1186/1471-2164-15-545. URL http://www.biomedcentral.com/1471-2164/15/545.

Hamblin, M. T. and Di Rienzo, A. Detection of the signature of natural selection in humans : evidence from the Duffy blood group locus. *American Journal of Human Genetics*, 66(5) :1669–1679, May 2000. ISSN 0002-9297. doi : 10.1086/302879.

Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C., and Wall, J. D. Genetic evidence for archaic admixture in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 108(37) :15123–15128, Sept. 2011. ISSN 1091-6490. doi : 10.1073/pnas.1109300108.

Han, Y., Gu, S., Oota, H., Osier, M. V., Pakstis, A. J., Speed, W. C., Kidd, J. R., and Kidd, K. K. Evidence of Positive Selection on a Class I ADH Locus. *The American Journal of Human Genetics*, 80(3) :441–456, Mar. 2007. ISSN 00029297. doi : 10.1086/512485. URL http://linkinghub.elsevier.com/retrieve/pii/S0002929707600937.

Hancock, A. M., Witonsky, D. B., Alkorta-Aranburu, G., Beall, C. M., Gebremedhin, A., Sukernik, R., Utermann, G., Pritchard, J. K., Coop, G., and Di Rienzo, A. Adaptations to climate-mediated selective pressures in humans. *PLoS genetics*, 7 (4) :e1001375, Apr. 2011. ISSN 1553-7404. doi : 10.1371/journal.pgen.1001375.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.-B., Gao, Y., Deconde, R., Chen, M., Rajapakse, I., Friend, S., Ideker, T., and Zhang, K. Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Molecular Cell*, 49(2) :359–367, Jan. 2013. ISSN

10972765. doi : 10.1016/j.molcel.2012.10.016. URL http://linkinghub.elsevier.com/retrieve/pii/S1097276512008933.

Hansen, K. H., Bracken, A. P., Pasini, D., Dietrich, N., Gehani, S. S., Monrad, A., Rappsilber, J., Lerdrup, M., and Helin, K. A model for transmission of the H3k27me3 epigenetic mark. *Nature Cell Biology*, 10(11) :1291–1300, Nov. 2008. ISSN 1465-7392, 1476-4679. doi : 10.1038/ncb1787. URL http://www.nature.com/doifinder/10.1038/ncb1787.

Harding, R. M., Healy, E., Ray, A. J., Ellis, N. S., Flanagan, N., Todd, C., Dixon, C., Sajantila, A., Jackson, I. J., Birch-Machin, M. A., and Rees, J. L. Evidence for variable selective pressures at MC1r. *American Journal of Human Genetics*, 66(4) : 1351–1361, Apr. 2000. ISSN 0002-9297. doi : 10.1086/302863.

Harris, K. Evidence for recent, population-specific evolution of the human mutation rate. *Proceedings of the National Academy of Sciences*, 112(11) :3439–3444, Mar. 2015. doi : 10.1073/pnas.1418652112. URL http://www.pnas.org/content/112/11/3439.abstract.

Hashimoto, H., Liu, Y., Upadhyay, A. K., Chang, Y., Howerton, S. B., Vertino, P. M., Zhang, X., and Cheng, X. Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Research*, 40(11) :4841–4849, June 2012. ISSN 0305-1048, 1362-4962. doi : 10.1093/nar/gks155. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gks155.

Hattori, N. and Ushijima, T. Compendium of aberrant DNA methylation and histone modifications in cancer. *Biochemical and Biophysical Research Communications*, 455(1-2) :3–9, Dec. 2014. ISSN 0006291X. doi : 10.1016/j.bbrc.2014.08.140. URL http://linkinghub.elsevier.com/retrieve/pii/S0006291X14015812.

He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., Sun, Y., Li, X., Dai, Q., Song, C.-X., Zhang, K., He, C., and Xu, G.-L. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science (New York, N.Y.)*, 333(6047) :1303–1307, Sept. 2011. ISSN 1095-9203. doi : 10.1126/science.1210944.

Heard, E. and Martienssen, R. Transgenerational Epigenetic Inheritance : Myths and Mechanisms. *Cell*, 157(1) :95–109, Mar. 2014. ISSN 00928674. doi : 10.1016/j.cell.2014.02.045. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867414002864.

Hedrick, P. W. and Thomson, G. EVIDENCE FOR BALANCING SELECTION AT HLA. *Genetics*, 104(3) :449–456, July 1983. URL http://www.genetics.org/content/104/3/449.abstract.

Heijmans, B. T., Tobi, E. W., Stein, A. D., Putter, H., Blauw, G. J., Susser, E. S., Slagboom, P. E., and Lumey, L. H. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proceedings of the National Academy of Sciences*, 105(44) :17046–17049, Nov. 2008. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.0806560105. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0806560105.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., and Ren, B. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3) :311–318, Mar. 2007. ISSN 1061-4036. doi : 10.1038/ng1966. URL http://www.nature.com/doifinder/10.1038/ng1966.

Hellenthal, G., Auton, A., and Falush, D. Inferring Human Colonization History Using a Copying Model. *PLoS Genetics*, 4(5) :e1000078, May 2008. ISSN 1553-7404. doi : 10.1371/journal.pgen.1000078. URL http://dx.plos.org/10.1371/journal.pgen.1000078.

Hermisson, J. and Pennings, P. S. Soft sweeps : molecular population genetics of adaptation from standing genetic variation. *Genetics*, 169(4) :2335–2352, Apr. 2005. ISSN 0016-6731. doi : 10.1534/genetics.104.036947.

Hernandez, R. D., Kelley, J. L., Elyashiv, E., Melton, S. C., Auton, A., McVean, G., 1000 Genomes Project, Sella, G., and Przeworski, M. Classic Selective Sweeps Were Rare in Recent Human Evolution. *Science*, 331(6019) :920–924, Feb. 2011. doi : 10.1126/science.1198878. URL http://www.sciencemag.org/content/331/6019/920.abstract.

Hewlett, B. S., editor. *Hunter-gatherers of the Congo Basin : cultures, histories and biology of African Pygmies*. Transaction Publishers, New Brunswick, NJ, 2014. ISBN 9781412853613.

Heyn, H. A symbiotic liaison between the genetic and epigenetic code. *Frontiers in Genetics*, 5, May 2014. ISSN 1664-8021. doi : 10.3389/fgene.2014.00113. URL http://journal.frontiersin.org/article/10.3389/fgene.2014.00113/abstract.

Heyn, H., Ferreira, H. J., Bassas, L., Bonache, S., Sayols, S., Sandoval, J., Esteller, M., and Larriba, S. Epigenetic disruption of the PIWI pathway in human spermatogenic disorders. *PloS One*, 7(10) :e47892, 2012. ISSN 1932-6203. doi : 10.1371/journal.pone.0047892.

Heyn, H., Moran, S., Hernando-Herraez, I., Sayols, S., Gomez, A., Sandoval, J., Monk, D., Hata, K., Marques-Bonet, T., Wang, L., and Esteller, M. DNA methylation contributes to natural human variation. *Genome Research*, 23(9) : 1363–1372, Sept. 2013. ISSN 1088-9051. doi : 10.1101/gr.154187.112. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.154187.112.

Hinch, A. G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C. D., Chen, G. K., Wang, K., Buxbaum, S. G., Akylbekova, E. L., Aldrich, M. C., Ambrosone, C. B., Amos, C., Bandera, E. V., Berndt, S. I., Bernstein, L., Blot, W. J., Bock, C. H., Boerwinkle, E., Cai, Q., Caporaso, N., Casey, G., Adrienne Cupples, L., Deming, S. L., Ryan Diver, W., Divers, J., Fornage, M., Gillanders, E. M., Glessner, J., Harris, C. C., Hu, J. J., Ingles, S. A., Isaacs, W., John, E. M., Linda Kao, W. H., Keating, B., Kittles, R. A., Kolonel, L. N., Larkin, E., Le Marchand, L., McNeill, L. H., Millikan, R. C., Murphy, Musani, S., Neslund-Dudas, C., Nyante, S.,

Papanicolaou, G. J., Press, M. F., Psaty, B. M., Reiner, A. P., Rich, S. S., Rodriguez-Gil, J. L., Rotter, J. I., Rybicki, B. A., Schwartz, A. G., Signorello, L. B., Spitz, M., Strom, S. S., Thun, M. J., Tucker, M. A., Wang, Z., Wiencke, J. K., Witte, J. S., Wrensch, M., Wu, X., Yamamura, Y., Zanetti, K. A., Zheng, W., Ziegler, R. G., Zhu, X., Redline, S., Hirschhorn, J. N., Henderson, B. E., Taylor Jr, H. A., Price, A. L., Hakonarson, H., Chanock, S. J., Haiman, C. A., Wilson, J. G., Reich, D., and Myers, S. R. The landscape of recombination in African Americans. *Nature*, 476(7359) : 170–175, July 2011. ISSN 0028-0836, 1476-4687. doi : 10.1038/nature10336. URL http://www.nature.com/doifinder/10.1038/nature10336.

Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A., and Cox, D. R. Whole-genome patterns of common DNA variation in three human populations. *Science (New York, N.Y.)*, 307(5712) :1072–1079, Feb. 2005. ISSN 1095-9203. doi : 10.1126/science.1105436.

Ho, J. W. K., Jung, Y. L., Liu, T., Alver, B. H., Lee, S., Ikegami, K., Sohn, K.-A., Minoda, A., Tolstorukov, M. Y., Appert, A., Parker, S. C. J., Gu, T., Kundaje, A., Riddle, N. C., Bishop, E., Egelhofer, T. A., Hu, S. S., Alekseyenko, A. A., Rechtsteiner, A., Asker, D., Belsky, J. A., Bowman, S. K., Chen, Q. B., Chen, R. A.-J., Day, D. S., Dong, Y., Dose, A. C., Duan, X., Epstein, C. B., Ercan, S., Feingold, E. A., Ferrari, F., Garrigues, J. M., Gehlenborg, N., Good, P. J., Haseley, P., He, D., Herrmann, M., Hoffman, M. M., Jeffers, T. E., Kharchenko, P. V., Kolasinska-Zwierz, P., Kotwaliwale, C. V., Kumar, N., Langley, S. A., Larschan, E. N., Latorre, I., Libbrecht, M. W., Lin, X., Park, R., Pazin, M. J., Pham, H. N., Plachetka, A., Qin, B., Schwartz, Y. B., Shoresh, N., Stempor, P., Vielle, A., Wang, C., Whittle, C. M., Xue, H., Kingston, R. E., Kim, J. H., Bernstein, B. E., Dernburg, A. F., Pirrotta, V., Kuroda, M. I., Noble, W. S., Tullius, T. D., Kellis, M., MacAlpine, D. M., Strome, S., Elgin, S. C. R., Liu, X. S., Lieb, J. D., Ahringer, J., Karpen, G. H., and Park, P. J. Comparative analysis of metazoan chromatin organization. *Nature*, 512(7515) :449–452, Aug. 2014. ISSN 0028-0836. URL http://dx.doi.org/10.1038/nature13415.

Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10) :R115, 2013. ISSN 1465-6914. doi : 10.1186/gb-2013-14-10-r115.

Horvath, S., Erhart, W., Brosch, M., Ammerpohl, O., von Schönfels, W., Ahrens, M., Heits, N., Bell, J. T., Tsai, P.-C., Spector, T. D., Deloukas, P., Siebert, R., Sipos, B., Becker, T., Röcken, C., Schafmayer, C., and Hampe, J. Obesity accelerates epigenetic aging of human liver. *Proceedings of the National Academy of Sciences of the United States of America*, 111(43) :15538–15543, Oct. 2014. ISSN 1091-6490. doi : 10.1073/pnas.1412759111.

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., and Kelsey, K. T. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, 13 :86, 2012. ISSN 1471-2105. doi : 10.1186/1471-2105-13-86.

Hu, S., Wan, J., Su, Y., Song, Q., Zeng, Y., Nguyen, H. N., Shin, J., Cox, E., Rho, H. S., Woodard, C., Xia, S., Liu, S., Lyu, H., Ming, G.-L., Wade, H., Song,

H., Qian, J., and Zhu, H. DNA methylation presents distinct binding sites for human transcription factors. *eLife*, 2, Sept. 2013. ISSN 2050-084X. doi : 10.7554/eLife.00726. URL http://elifesciences.org/lookup/doi/10.7554/eLife.00726.

Hudson, R. R., Kreitman, M., and Aguadé, M. A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics*, 116(1) :153–159, May 1987. URL http://www.genetics.org/content/116/1/153.abstract.

Hudson, R. R., Slatkin, M., and Maddison, W. P. Estimation of levels of gene flow from DNA sequence data. *Genetics*, 132(2) :583–589, Oct. 1992. ISSN 0016-6731.

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., Wang, B., Ou, X., Huasang, Luosang, J., Cuo, Z. X. P., Li, K., Gao, G., Yin, Y., Wang, W., Zhang, X., Xu, X., Yang, H., Li, Y., Wang, J., Wang, J., and Nielsen, R. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, 512(7513) :194–197, July 2014. ISSN 0028-0836, 1476-4687. doi : 10.1038/nature13408. URL http://www.nature.com/doifinder/10.1038/nature13408.

Hurles, M. E., Dermitzakis, E. T., and Tyler-Smith, C. The functional impact of structural variation in humans. *Trends in genetics : TIG*, 24(5) :238–245, May 2008. ISSN 0168-9525. doi : 10.1016/j.tig.2008.03.001.

Hutchison, D. W. and Templeton, A. R. Correlation of Pairwise Genetic and Geographic Distance Measures : Inferring the Relative Influences of Gene Flow and Drift on the Distribution of Genetic Variability. *Evolution*, 53(6) :1898, Dec. 1999. ISSN 00143820. doi : 10.2307/2640449. URL http://www.jstor.org/stable/2640449?origin=crossref.

Idaghdour, Y., Storey, J. D., Jadallah, S. J., and Gibson, G. A Genome-Wide Gene Expression Signature of Environmental Geography in Leukocytes of Moroccan Amazighs. *PLoS Genetics*, 4(4) :e1000052, Apr. 2008. ISSN 1553-7404. doi : 10.1371/journal.pgen.1000052. URL http://dx.plos.org/10.1371/journal.pgen.1000052.

Ingman, M., Kaessmann, H., Paabo, S., and Gyllensten, U. Mitochondrial genome variation and the origin of modern humans. *Nature*, 408(6813) :708–713, Dec. 2000. ISSN 0028-0836. doi : 10.1038/35047064. URL http://dx.doi.org/10.1038/35047064.

Innan, H. and Kim, Y. Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. *Genetics*, 179(3) : 1713–1720, July 2008. ISSN 0016-6731. doi : 10.1534/genetics.108.086835.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011) :931–945, Oct. 2004. ISSN 0028-0836. doi : 10.1038/nature03001. URL http://dx.doi.org/10.1038/nature03001.

Itan, Y., Jones, B., Ingram, C., Swallow, D., and Thomas, M. A worldwide correlation of lactase persistence phenotype and genotypes. *BMC Evolutionary Biology*, 10 (1) :36, 2010. URL http://www.biomedcentral.com/1471-2148/10/36.

Ito, S., Shen, L., Dai, Q., Wu, S. C., Collins, L. B., Swenberg, J. A., He, C., and Zhang, Y. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science (New York, N.Y.)*, 333(6047) :1300–1303, Sept. 2011. ISSN 1095-9203. doi : 10.1126/science.1210597.

Izagirre, N., García, I., Junquera, C., de la Rúa, C., and Alonso, S. A Scan for Signatures of Positive Selection in Candidate Loci for Skin Pigmentation in Humans. *Molecular Biology and Evolution*, 23(9) :1697–1706, Sept. 2006. doi : 10.1093/molbev/msl030. URL http://mbe.oxfordjournals.org/content/23/9/1697.abstract.

Jabbari, K. and Bernardi, G. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene*, 333 :143–149, May 2004. ISSN 03781119. doi : 10.1016/j.gene.2004.02.043. URL http://linkinghub.elsevier.com/retrieve/pii/S0378111904000836.

Jablonka, E. and Raz, G. Transgenerational Epigenetic Inheritance : Prevalence, Mechanisms, and Implications for the Study of Heredity and Evolution. *The Quarterly Review of Biology*, 84(2) :131–176, June 2009. ISSN 0033-5770, 1539-7718. doi : 10.1086/598822. URL http://www.jstor.org/stable/10.1086/598822.

Jablonski, N. G. and Chaplin, G. The evolution of human skin coloration. *Journal of Human Evolution*, 39(1) :57–106, July 2000. ISSN 0047-2484. doi : 10.1006/jhev.2000.0403.

Jaenisch, R. and Bird, A. Epigenetic regulation of gene expression : how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33(3s) :245–254, Mar. 2003. ISSN 10614036. doi : 10.1038/ng1089. URL http://www.nature.com/doifinder/10.1038/ng1089.

Jarvis, J. P., Scheinfeldt, L. B., Soi, S., Lambert, C., Omberg, L., Ferwerda, B., Froment, A., Bodo, J.-M., Beggs, W., Hoffman, G., Mezey, J., and Tishkoff, S. A. Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS genetics*, 8(4) :e1002641, 2012. ISSN 1553-7404. doi : 10.1371/journal.pgen.1002641.

Jiang, N., Wang, L., Chen, J., Wang, L., Leach, L., and Luo, Z. Conserved and Divergent Patterns of DNA Methylation in Higher Vertebrates. *Genome Biology and Evolution*, 6(11) :2998–3014, Nov. 2014. ISSN 1759-6653. doi : 10.1093/gbe/evu238. URL http://gbe.oxfordjournals.org/cgi/doi/10.1093/gbe/evu238.

Jin, W., Xu, S., Wang, H., Yu, Y., Shen, Y., Wu, B., and Jin, L. Genome-wide detection of natural selection in African Americans pre-and post-admixture. *Genome Research*, Nov. 2011. doi : 10.1101/gr.124784.111. URL http://genome.cshlp.org/content/early/2011/11/29/gr.124784.111.abstract.

Jones, P. A. Functions of DNA methylation : islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7) :484–492, May 2012. ISSN 1471-0056, 1471-0064. doi : 10.1038/nrg3230. URL http://www.nature.com/doifinder/10.1038/nrg3230.

Joubert, B. R., Håberg, S. E., Bell, D. A., Nilsen, R. M., Vollset, S. E., Midttun, O., Ueland, P. M., Wu, M. C., Nystad, W., Peddada, S. D., and London, S. J. Maternal smoking and DNA methylation in newborns : in utero effect or epigenetic inheritance ? *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, 23(6) :1007–1017, June 2014. ISSN 1538-7755. doi : 10.1158/1055-9965.EPI-13-1256.

Kamberov, Y., Wang, S., Tan, J., Gerbault, P., Wark, A., Tan, L., Yang, Y., Li, S., Tang, K., Chen, H., Powell, A., Itan, Y., Fuller, D., Lohmueller, J., Mao, J., Schachar, A., Paymer, M., Hostetter, E., Byrne, E., Burnett, M., McMahon, A., Thomas, M., Lieberman, D., Jin, L., Tabin, C., Morgan, B., and Sabeti, P. Modeling Recent Human Evolution in Mice by Expression of a Selected EDAR Variant. *Cell*, 152 (4) :691–702, Feb. 2013. ISSN 00928674. doi : 10.1016/j.cell.2013.01.016. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867413000676.

Kaplow, I. M., MacIsaac, J. L., Mah, S. M., McEwen, L. M., Kobor, M. S., and Fraser, H. B. A pooling-based approach to mapping genetic variants associated with DNA methylation. *Genome Research*, page gr.183749.114, Apr. 2015. ISSN 1088-9051. doi : 10.1101/gr.183749.114. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.183749.114.

Karlsson, E. K., Kwiatkowski, D. P., and Sabeti, P. C. Natural selection and infectious disease in human populations. *Nat Rev Genet*, 15(6) :379–393, June 2014. ISSN 1471-0056. URL http://dx.doi.org/10.1038/nrg3734.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., Hong, M. Y., Karczewski, K. J., Huber, W., Weissman, S. M., Gerstein, M. B., Korbel, J. O., and Snyder, M. Variation in Transcription Factor Binding Among Humans. *Science*, 328(5975) : 232–235, Apr. 2010. ISSN 0036-8075, 1095-9203. doi : 10.1126/science.1183621. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1183621.

Katz, D. H. and Benacerraf, B., editors. *The Role of products of the histocompatibility gene complex in immune responses : [proceedings of an international conference held at Brook Lodge, Augusta, Michigan, November 3-7, 1975].* Academic Press, New York, 1976. ISBN 012401660X.

Kawahara, M., Wu, Q., Takahashi, N., Morita, S., Yamada, K., Ito, M., Ferguson-Smith, A. C., and Kono, T. High-frequency generation of viable mice from engineered bi-maternal embryos. *Nature Biotechnology*, 25(9) :1045–1050, Sept. 2007. ISSN 1087-0156. doi : 10.1038/nbt1331.

Kelley, J. L. and Swanson, W. J. Positive Selection in the Human Genome : From Genome Scans to Biological Significance. *Annual Review of Genomics and Human*

Genetics, 9(1) :143–160, Sept. 2008. ISSN 1527-8204, 1545-293X. doi : 10.1146/annurev.genom.9.081307.164411. URL http://www.annualreviews.org/doi/abs/10.1146/annurev.genom.9.081307.164411.

Kelley, J. L., Madeoy, J., Calhoun, J. C., Swanson, W., and Akey, J. M. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Research*, 16(8) :980–989, Aug. 2006. doi : 10.1101/gr.5157306. URL http://genome.cshlp.org/content/16/8/980.abstract.

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., Ward, L. D., Birney, E., Crawford, G. E., Dekker, J., Dunham, I., Elnitski, L. L., Farnham, P. J., Feingold, E. A., Gerstein, M., Giddings, M. C., Gilbert, D. M., Gingeras, T. R., Green, E. D., Guigo, R., Hubbard, T., Kent, J., Lieb, J. D., Myers, R. M., Pazin, M. J., Ren, B., Stamatoyannopoulos, J. A., Weng, Z., White, K. P., and Hardison, R. C. Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences*, 111(17) :6131–6138, Apr. 2014. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.1318948111. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1318948111.

Kerkel, K., Spadola, A., Yuan, E., Kosek, J., Jiang, L., Hod, E., Li, K., Murty, V. V., Schupf, N., Vilain, E., Morris, M., Haghighi, F., and Tycko, B. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nature Genetics*, 40(7) :904–908, July 2008. ISSN 1061-4036. doi : 10.1038/ng.174. URL http://www.nature.com/doifinder/10.1038/ng.174.

Key, F. M., Teixeira, J. C., de Filippo, C., and Andrés, A. M. Advantageous diversity maintained by balancing selection in humans. *Current Opinion in Genetics & Development*, 29 :45–51, Dec. 2014. ISSN 0959437X. doi : 10.1016/j.gde.2014.08.001. URL http://linkinghub.elsevier.com/retrieve/pii/S0959437X14000823.

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B. E., van Oudenaarden, A., Regev, A., Lander, E. S., and Rinn, J. L. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences*, 106(28) :11667–11672, July 2009. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.0904715106. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0904715106.

Khorasanizadeh, S. The nucleosome : from genomic organization to genomic regulation. *Cell*, 116(2) :259–272, Jan. 2004. ISSN 0092-8674.

Khrameeva, E. E., Bozek, K., He, L., Yan, Z., Jiang, X., Wei, Y., Tang, K., Gelfand, M. S., Prufer, K., Kelso, J., Paabo, S., Giavalisco, P., Lachmann, M., and Khaitovich, P. Neanderthal ancestry drives evolution of lipid catabolism in contemporary Europeans. *Nature Communications*, 5, Apr. 2014. ISSN 2041-1723. doi : 10.1038/ncomms4584. URL http://www.nature.com/doifinder/10.1038/ncomms4584.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tuzun, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., McCarroll, S. A., Altshuler, D. A., Peiffer, D. A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D. A., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R., and Eichler, E. E. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453 (7191) :56–64, May 2008. ISSN 0028-0836. doi : 10.1038/nature06862. URL http://dx.doi.org/10.1038/nature06862.

Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D., and Ren, B. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052) :876–880, Aug. 2005. ISSN 0028-0836. doi : 10.1038/nature03877. URL http://dx.doi.org/10.1038/nature03877.

Kimura, M. and Weiss, G. H. The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. *Genetics*, 49(4) :561–576, Apr. 1964. ISSN 0016-6731.

Kimura, R., Fujimoto, A., Tokunaga, K., and Ohashi, J. A Practical Genome Scan for Population-Specific Strong Selective Sweeps That Have Reached Fixation. *PLoS ONE*, 2(3) :e286, Mar. 2007. ISSN 1932-6203. doi : 10.1371/journal.pone.0000286. URL http://dx.plos.org/10.1371/journal.pone.0000286.

King, M. and Wilson, A. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184) :107–116, Apr. 1975. ISSN 0036-8075, 1095-9203. doi : 10.1126/science.1090005. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1090005.

Klein, J., Satta, Y., O'hUigin, C., and Takahata, N. The molecular descent of the major histocompatibility complex. *Annual Review of Immunology*, 11 :269–295, 1993. ISSN 0732-0582. doi : 10.1146/annurev.iy.11.040193.001413.

Klimentidis, Y. C., Abrams, M., Wang, J., Fernandez, J. R., and Allison, D. B. Natural selection at genomic regions associated with obesity and type-2 diabetes : East Asians and sub-Saharan Africans exhibit high levels of differentiation at type-2 diabetes regions. *Human Genetics*, 129(4) :407–418, Apr. 2011. ISSN 0340-6717, 1432-1203. doi : 10.1007/s00439-010-0935-z. URL http://link.springer.com/10.1007/s00439-010-0935-z.

Klironomos, F. D., Berg, J., and Collins, S. How epigenetic mutations can affect genetic evolution : Model and mechanism : Problems &amp ; Paradigms. *BioEssays*, 35(6) :571–578, June 2013. ISSN 02659247. doi : 10.1002/bies.201200169. URL http://doi.wiley.com/10.1002/bies.201200169.

Klopfstein, S., Currat, M., and Excoffier, L. The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*, 23(3) :482–490, Mar. 2006. ISSN 0737-4038. doi : 10.1093/molbev/msj057.

Klose, R. J. and Bird, A. P. Genomic DNA methylation : the mark and its mediators. *Trends in Biochemical Sciences*, 31(2) :89–97, Feb. 2006. ISSN 0968-0004. doi : 10.1016/j.tibs.2005.12.008.

Koestler, D. C., Christensen, B., Karagas, M. R., Marsit, C. J., Langevin, S. M., Kelsey, K. T., Wiencke, J. K., and Houseman, E. A. Blood-based profiles of DNA methylation predict the underlying distribution of cell types : a validation analysis. *Epigenetics : official journal of the DNA Methylation Society*, 8(8) :816–826, Aug. 2013. ISSN 1559-2308. doi : 10.4161/epi.25430.

Kohli, R. M. and Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature*, 502(7472) :472–479, Oct. 2013. ISSN 0028-0836, 1476-4687. doi : 10.1038/nature12750. URL http://www.nature.com/doifinder/10.1038/nature12750.

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., and Stefansson, K. A high-resolution recombination map of the human genome. *Nature Genetics*, 31(3) :241–247, July 2002. ISSN 1061-4036. doi : 10.1038/ng917.

Kornberg, R. D. Chromatin structure : a repeating unit of histones and DNA. *Science (New York, N.Y.)*, 184(4139) :868–871, May 1974. ISSN 0036-8075.

Kouzarides, T. Chromatin Modifications and Their Function. *Cell*, 128(4) :693–705, Feb. 2007. ISSN 00928674. doi : 10.1016/j.cell.2007.02.005. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867407001845.

Kudaravalli, S., Veyrieras, J.-B., Stranger, B. E., Dermitzakis, E. T., and Pritchard, J. K. Gene Expression Levels Are a Target of Recent Natural Selection in the Human Genome. *Molecular Biology and Evolution*, 26(3) :649 –658, Mar. 2009. doi : 10.1093/molbev/msn289. URL http://mbe.oxfordjournals.org/content/26/3/649.abstract.

Labrador, M. and Corces, V. G. Setting the Boundaries of Chromatin Domains and Nuclear Organization. *Cell*, 111(2) :151–154, Oct. 2002. ISSN 00928674. doi : 10.1016/S0092-8674(02)01004-8. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867402010048.

Lachance, J. and Tishkoff, S. A. Population Genomics of Human Adaptation. *Annual Review of Ecology, Evolution, and Systematics*, 44 :123–143, Nov. 2013. ISSN 1543-592X. doi : 10.1146/annurev-ecolsys-110512-135833.

Lachance, J., Vernot, B., Elbers, C. C., Ferwerda, B., Froment, A., Bodo, J.-M., Lema, G., Fu, W., Nyambo, T. B., Rebbeck, T. R., Zhang, K., Akey, J. M., and Tishkoff, S. A. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell*, 150(3) :457–469, Aug. 2012. ISSN 1097-4172. doi : 10.1016/j.cell.2012.07.009.

Laird, P. W. Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, 11(3) :191, Mar. 2010. ISSN 1471-0056, 1471-0064. doi : 10.1038/nrg2732. URL http://www.nature.com/doifinder/10.1038/nrg2732.

Lam, L. L., Emberly, E., Fraser, H. B., Neumann, S. M., Chen, E., Miller, G. E., and Kobor, M. S. Factors underlying variable DNA methylation in a human community cohort. *Proceedings of the National Academy of Sciences*, 109(Supplement 2) : 17253–17260, Oct. 2012. doi : 10.1073/pnas.1121249109. URL http://www.pnas.org/content/109/Supplement_2/17253.abstract.

LaRocca, J., Binder, A. M., McElrath, T. F., and Michels, K. B. The impact of first trimester phthalate and phenol exposure on IGF2/H19 genomic imprinting and birth outcomes. *Environmental Research*, 133 :396–406, Aug. 2014. ISSN 1096-0953. doi : 10.1016/j.envres.2014.04.032.

Laval, G., Patin, E., Barreiro, L. B., and Quintana-Murci, L. Formulating a Historical and Demographic Model of Recent Human Evolution Based on Resequencing Data from Noncoding Regions. *PLoS ONE*, 5(4) :e10284, 2010. doi : 10.1371/journal.pone.0010284. URL http://dx.doi.org/10.1371%2Fjournal.pone.0010284.

Lesecque, Y., Glémin, S., Lartillot, N., Mouchiroud, D., and Duret, L. The Red Queen Model of Recombination Hotspots Evolution in the Light of Archaic and Modern Human Genomes. *PLoS Genetics*, 10(11) :e1004790, Nov. 2014. ISSN 1553-7404. doi : 10.1371/journal.pgen.1004790. URL http://dx.plos.org/10.1371/journal.pgen.1004790.

Lewontin, R. C. and Krakauer, J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, 74(1) :175–195, May 1973. ISSN 0016-6731.

Li, E., Bestor, T. H., and Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, 69(6) :915–926, June 1992. ISSN 00928674. doi : 10.1016/0092-8674(92)90611-F. URL http://linkinghub.elsevier.com/retrieve/pii/009286749290611F.

Li, E., Beard, C., and Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature*, 366(6453) :362–365, Dec. 1993. ISSN 0028-0836. doi : 10.1038/366362a0. URL http://www.nature.com/doifinder/10.1038/366362a0.

Li, H. A new test for detecting recent positive selection that is free from the confounding impacts of demography. *Molecular Biology and Evolution*, 28(1) : 365–375, Jan. 2011. ISSN 1537-1719. doi : 10.1093/molbev/msq211.

Li, M. J., Wang, L. Y., Xia, Z., Wong, M. P., Sham, P. C., and Wang, J. dbPSHP : a database of recent positive selection across human populations. *Nucleic Acids Research*, 42(Database issue) :D910–916, Jan. 2014. ISSN 1362-4962. doi : 10.1093/nar/gkt1052.

Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., and Hemminki, K. Environmental and Heritable Factors in the Causation of Cancer — Analyses of Cohorts of Twins from Sweden, Denmark, and Finland. *New England Journal of Medicine*, 343(2) :78–85, July 2000. ISSN 0028-4793, 1533-4406. doi : 10.1056/NEJM200007133430201. URL http://www.nejm.org/doi/abs/10.1056/NEJM200007133430201.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271) :315–322, Nov. 2009. ISSN 1476-4687. doi : 10.1038/nature08514.

Lohmueller, K. E. The distribution of deleterious genetic variation in human populations. *Current Opinion in Genetics & Development*, 29 :139–146, Dec. 2014. ISSN 0959437X. doi : 10.1016/j.gde.2014.09.005. URL http://linkinghub.elsevier.com/retrieve/pii/S0959437X14001002.

Lohmueller, K. E., Indap, A. R., Schmidt, S., Boyko, A. R., Hernandez, R. D., Hubisz, M. J., Sninsky, J. J., White, T. J., Sunyaev, S. R., Nielsen, R., Clark, A. G., and Bustamante, C. D. Proportionally more deleterious genetic variation in European than in African populations. *Nature*, 451(7181) :994–997, Feb. 2008. ISSN 0028-0836, 1476-4687. doi : 10.1038/nature06611. URL http://www.nature.com/doifinder/10.1038/nature06611.

Lohmueller, K. E., Albrechtsen, A., Li, Y., Kim, S. Y., Korneliussen, T., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Feder, A. F., Grarup, N., Jørgensen, T., Jiang, T., Witte, D. R., Sandbæk, A., Hellmann, I., Lauritzen, T., Hansen, T., Pedersen, O., Wang, J., and Nielsen, R. Natural Selection Affects Multiple Aspects of Genetic Variation at Putatively Neutral Sites across the Human Genome. *PLoS Genetics*, 7 (10) :e1002326, Oct. 2011. ISSN 1553-7404. doi : 10.1371/journal.pgen.1002326. URL http://dx.plos.org/10.1371/journal.pgen.1002326.

Louicharoen, C., Patin, E., Paul, R., Nuchprayoon, I., Witoonpanich, B., Peerapittayamongkol, C., Casademont, I., Sura, T., Laird, N. M., Singhasivanon, P., Quintana-Murci, L., and Sakuntabhai, A. Positively Selected G6pd-Mahidol Mutation Reduces Plasmodium vivax Density in Southeast Asians. *Science*, 326 (5959) :1546 –1549, Dec. 2009. doi : 10.1126/science.1178849. URL http://www.sciencemag.org/content/326/5959/1546.abstract.

Luca, F., Bubba, G., Basile, M., Brdicka, R., Michalodimitrakis, E., Rickards, O., Vershubsky, G., Quintana-Murci, L., Kozlov, A. I., and Novelletto, A. Multiple advantageous amino acid variants in the NAT2 gene in human populations. *PloS One*, 3(9) :e3136, 2008. ISSN 1932-6203. doi : 10.1371/journal.pone.0003136.

Luca, F., Perry, G. H., and Di Rienzo, A. Evolutionary adaptations to dietary changes. *Annual Review of Nutrition*, 30 :291–314, Aug. 2010. ISSN 1545-4312. doi : 10.1146/annurev-nutr-080508-141048.

Luisi, P., Alvarez-Ponce, D., Pybus, M., Fares, M. A., Bertranpetit, J., and Laayouni, H. Recent Positive Selection Has Acted On Genes Encoding Proteins With More Interactions Within the Whole Human Interactome. *Genome Biology and Evolution*, Apr. 2015. ISSN 1759-6653. doi : 10.1093/gbe/evv055.

Lupski, J. R. Genomic rearrangements and sporadic disease. *Nat Genet*, June 2007.

Macaulay, V. Single, Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes. *Science*, 308(5724) :1034–1036, May 2005. ISSN 0036-8075, 1095-9203. doi : 10.1126/science.1109792. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1109792.

MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., and Scherer, S. W. The Database of Genomic Variants : a curated collection of structural variation in the human genome. *Nucleic Acids Research*, 42(D1) :D986–D992, Jan. 2014. doi : 10.1093/nar/gkt958. URL http://nar.oxfordjournals.org/content/42/D1/D986.abstract.

Majnik, A. V. and Lane, R. H. Epigenetics : where environment, society and genetics meet. *Epigenomics*, 6(1) :1–4, Feb. 2014. ISSN 1750-192X. doi : 10.2217/epi.13.83.

Manikkam, M., Guerrero-Bosagna, C., Tracey, R., Haque, M. M., and Skinner, M. K. Transgenerational Actions of Environmental Compounds on Reproductive Disease and Identification of Epigenetic Biomarkers of Ancestral Exposures. *PLoS ONE*, 7(2) :e31901, Feb. 2012. ISSN 1932-6203. doi : 10.1371/journal.pone.0031901. URL http://dx.plos.org/10.1371/journal.pone.0031901.

Mann, J. R. and Lovell-Badge, R. H. Inviability of parthenogenones is determined by pronuclei, not egg cytoplasm. *Nature*, 310(5972) :66–67, July 1984. ISSN 0028-0836.

Manry, J., Laval, G., Patin, E., Fornarino, S., Itan, Y., Fumagalli, M., Sironi, M., Tichit, M., Bouchier, C., Casanova, J.-L., Barreiro, L. B., and Quintana-Murci, L. Evolutionary genetic dissection of human interferons. *The Journal of Experimental Medicine*, 208(13) :2747–2759, Dec. 2011. ISSN 1540-9538. doi : 10.1084/jem.20111680.

Mayr, E. *Animal species and evolution.* Havard University Press, 1963.

Mazzio, E. A. and Soliman, K. F. A. Basic concepts of epigenetics : impact of environmental signals on gene expression. *Epigenetics : official journal of the DNA Methylation Society*, 7(2) :119–130, Feb. 2012. ISSN 1559-2308. doi : 10.4161/epi.7.2.18764.

McDonald, J. H. and Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328) :652–654, June 1991. ISSN 0028-0836. doi : 10.1038/351652a0.

McDougall, I., Brown, F. H., and Fleagle, J. G. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature*, 433(7027) :733–736, Feb. 2005. ISSN 1476-4687. doi : 10.1038/nature03258.

McGrath, J. and Solter, D. Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell*, 37(1) :179–183, May 1984. ISSN 0092-8674.

McRae, A. F., Powell, J. E., Henders, A. K., Bowdler, L., Hemani, G., Shah, S., Painter, J. N., Martin, N. G., Visscher, P. M., and Montgomery, G. W. Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome Biology*, 15(5) :R73, 2014. ISSN 1465-6906. doi : 10.1186/gb-2014-15-5-r73. URL http://genomebiology.com/2014/15/5/R73.

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. *Science*, 304(5670) :581–584, Apr. 2004. doi : 10.1126/science.1092500. URL http://www.sciencemag.org/content/304/5670/581.abstract.

Meissner, A. Epigenetic modifications in pluripotent and differentiated cells. *Nat Biotech*, 28(10) :1079–1088, Oct. 2010. ISSN 1087-0156. URL http://dx.doi.org/10.1038/nbt.1684.

Mellars, P. Going east : new genetic and archaeological perspectives on the modern human colonization of Eurasia. *Science (New York, N.Y.)*, 313(5788) :796–800, Aug. 2006. ISSN 1095-9203. doi : 10.1126/science.1128402.

Mendez, F. L., Watkins, J. C., and Hammer, M. F. A haplotype at STAT2 Introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. *American Journal of Human Genetics*, 91(2) :265–274, Aug. 2012. ISSN 1537-6605. doi : 10.1016/j.ajhg.2012.06.015.

Mendez, F. L., Watkins, J. C., and Hammer, M. F. Neandertal Origin of Genetic Variation at the Cluster of OAS Immunity Genes. *Molecular Biology and Evolution*, 30(4) :798–801, Apr. 2013. ISSN 0737-4038, 1537-1719. doi : 10.1093/molbev/mst004. URL http://mbe.oxfordjournals.org/cgi/doi/10.1093/molbev/mst004.

Mendizabal, I., Keller, T. E., Zeng, J., and Yi, S. V. Epigenetics and Evolution. *Integrative and Comparative Biology*, 54(1) :31–42, July 2014. ISSN 1540-7063, 1557-7023. doi : 10.1093/icb/icu040. URL http://icb.oxfordjournals.org/cgi/doi/10.1093/icb/icu040.

Messer, P. W. and Petrov, D. A. Frequent adaptation and the McDonald–Kreitman test. *Proceedings of the National Academy of Sciences*, 110(21) :8615–8620, May 2013a. doi : 10.1073/pnas.1220835110. URL http://www.pnas.org/content/110/21/8615.abstract.

Messer, P. W. and Petrov, D. A. Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution*, 28(11) :659–669, Nov. 2013b. ISSN 01695347. doi : 10.1016/j.tree.2013.08.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0169534713002073.

Moen, E. L., Zhang, X., Mu, W., Delaney, S. M., Wing, C., McQuade, J., Myers, J., Godley, L. A., Dolan, M. E., and Zhang, W. Genome-wide variation of cytosine

modifications between European and African populations and the implications for complex traits. *Genetics*, 194(4) :987–996, Aug. 2013. ISSN 1943-2631. doi : 10.1534/genetics.113.151381.

Morange, M. The relations between genetics and epigenetics : a historical point of view. *Annals of the New York Academy of Sciences*, 981 :50–60, Dec. 2002. ISSN 0077-8923.

Murgatroyd, C., Patchev, A. V., Wu, Y., Micale, V., Bockmühl, Y., Fischer, D., Holsboer, F., Wotjak, C. T., Almeida, O. F. X., and Spengler, D. Dynamic DNA methylation programs persistent adverse effects of early-life stress. *Nature Neuroscience*, 12(12) :1559–1566, Dec. 2009. ISSN 1097-6256, 1546-1726. doi : 10.1038/nn.2436. URL http://www.nature.com/doifinder/10.1038/nn.2436.

Murgatroyd, C., Wu, Y., Bockmühl, Y., and Spengler, D. The Janus face of DNA methylation in aging. *Aging*, 2(2) :107–110, 2010. ISSN 1945-4589.

Murphy, T. M. and Mill, J. Epigenetics in health and disease : heralding the EWAS era. *Lancet*, 383(9933) :1952–1954, June 2014. ISSN 1474-547X. doi : 10.1016/S0140-6736(14)60269-5.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. *Science*, 310 (5746) :321–324, Oct. 2005. doi : 10.1126/science.1117196. URL http://www.sciencemag.org/content/310/5746/321.abstract.

Myers, S., Freeman, C., Auton, A., Donnelly, P., and McVean, G. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet*, 40(9) :1124–1129, Sept. 2008. ISSN 1061-4036. doi : 10.1038/ng.213. URL http://dx.doi.org/10.1038/ng.213.

Nachman, M. W. and Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1) :297–304, Sept. 2000. ISSN 0016-6731.

Neel, J. V. Diabetes mellitus : a "thrifty" genotype rendered detrimental by "progress" ? *American Journal of Human Genetics*, 14 :353–362, Dec. 1962. ISSN 0002-9297.

Nei, M. *Molecular Evolutionary Genetics*. Columbia University Press, 1987. ISBN 9780231063210. URL http://books.google.fr/books?id=UhRSsLkjxDgC.

Nielsen, C. H., Larsen, A., and Nielsen, A. L. DNA methylation alterations in response to prenatal exposure of maternal cigarette smoking : A persistent epigenetic impact on health from maternal lifestyle ? *Archives of Toxicology*, Dec. 2014. ISSN 0340-5761, 1432-0738. doi : 10.1007/s00204-014-1426-0. URL http://link.springer.com/10.1007/s00204-014-1426-0.

Nielsen, R. Molecular signatures of natural selection. *Annual Review of Genetics*, 39 : 197–218, 2005. ISSN 0066-4197. doi : 10.1146/annurev.genet.39.073003.112420.

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., Fledel-Alon, A., Tanenbaum, D. M., Civello, D., White, T. J., J Sninsky, J., Adams, M. D., and Cargill, M. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS biology*, 3(6) :e170, June 2005a. ISSN 1545-7885. doi : 10.1371/journal.pbio.0030170.

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. Genomic scans for selective sweeps using SNP data. *Genome Research*, 15(11) : 1566–1575, Nov. 2005b. doi : 10.1101/gr.4252305. URL http://genome.cshlp.org/content/15/11/1566.abstract.

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., and Clark, A. G. Recent and ongoing selection in the human genome. *Nat Rev Genet*, 8(11) :857–868, Nov. 2007. ISSN 1471-0056. doi : 10.1038/nrg2187. URL http://dx.doi.org/10.1038/nrg2187.

Nielsen, R., Hubisz, M. J., Hellmann, I., Torgerson, D., Andrés, A. M., Albrechtsen, A., Gutenkunst, R., Adams, M. D., Cargill, M., Boyko, A., Indap, A., Bustamante, C. D., and Clark, A. G. Darwinian and demographic forces affecting human protein coding genes. *Genome Research*, 19(5) :838–849, May 2009. ISSN 1088-9051. doi : 10.1101/gr.088336.108.

Novembre, J. and Di Rienzo, A. Spatial patterns of variation due to natural selection in humans. *Nature Reviews. Genetics*, 10(11) :745–755, Nov. 2009. ISSN 1471-0064. doi : 10.1038/nrg2632.

Okano, M., Bell, D. W., Haber, D. A., and Li, E. DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. *Cell*, 99(3) :247–257, Oct. 1999. ISSN 00928674. doi : 10.1016/S0092-8674(00)81656-6. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867400816566.

Oleksyk, T. K., Zhao, K., De La Vega, F. M., Gilbert, D. A., O'Brien, S. J., and Smith, M. W. Identifying Selected Regions from Heterozygosity and Divergence Using a Light-Coverage Genomic Dataset from Two Human Populations. *PLoS ONE*, 3 (3) :e1712, Mar. 2008. ISSN 1932-6203. doi : 10.1371/journal.pone.0001712. URL http://dx.plos.org/10.1371/journal.pone.0001712.

Ollikainen, M., Smith, K. R., Joo, E. J.-H., Ng, H. K., Andronikos, R., Novakovic, B., Abdul Aziz, N. K., Carlin, J. B., Morley, R., Saffery, R., and Craig, J. M. DNA methylation analysis of multiple tissues from newborn twins reveals both genetic and intrauterine components to variation in the human neonatal epigenome. *Human Molecular Genetics*, 19(21) :4176–4188, Nov. 2010. ISSN 0964-6906, 1460-2083. doi : 10.1093/hmg/ddq336. URL http://www.hmg.oxfordjournals.org/cgi/doi/10.1093/hmg/ddq336.

Oslisly, R., White, L., Bentaleb, I., Favier, C., Fontugne, M., Gillet, J.-F., and Sebag, D. Climatic and cultural changes in the west Congo Basin forests over the past 5000 years. *Philosophical Transactions of the Royal Society of London. Series B,*

Biological Sciences, 368(1625) :20120304, 2013. ISSN 1471-2970. doi : 10.1098/rstb.2012.0304.

Pai, A. A., Bell, J. T., Marioni, J. C., Pritchard, J. K., and Gilad, Y. A Genome-Wide Study of DNA Methylation Patterns and Gene Expression Levels in Multiple Human and Chimpanzee Tissues. *PLoS Genetics*, 7(2) :e1001316, Feb. 2011. ISSN 1553-7404. doi : 10.1371/journal.pgen.1001316. URL http://dx.plos.org/10.1371/journal.pgen.1001316.

Pai, A. A., Pritchard, J. K., and Gilad, Y. The genetic and mechanistic basis for variation in gene regulation. *PLoS genetics*, 11(1) :e1004857, Jan. 2015. ISSN 1553-7404. doi : 10.1371/journal.pgen.1004857.

Patin, E., Barreiro, L. B., Sabeti, P. C., Austerlitz, F., Luca, F., Sajantila, A., Behar, D. M., Semino, O., Sakuntabhai, A., Guiso, N., Gicquel, B., McElreavey, K., Harding, R. M., Heyer, E., and Quintana-Murci, L. Deciphering the ancient and complex evolutionary history of human arylamine N-acetyltransferase genes. *American Journal of Human Genetics*, 78(3) :423–436, Mar. 2006. ISSN 0002-9297. doi : 10.1086/500614.

Patin, E., Laval, G., Barreiro, L. B., Salas, A., Semino, O., Santachiara-Benerecetti, S., Kidd, K. K., Kidd, J. R., Van der Veen, L., Hombert, J.-M., Gessain, A., Froment, A., Bahuchet, S., Heyer, E., and Quintana-Murci, L. Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS genetics*, 5(4) :e1000448, Apr. 2009. ISSN 1553-7404. doi : 10.1371/journal.pgen.1000448.

Patin, E., Siddle, K. J., Laval, G., Quach, H., Harmant, C., Becker, N., Froment, A., Régnault, B., Lemée, L., Gravel, S., Hombert, J.-M., Van der Veen, L., Dominy, N. J., Perry, G. H., Barreiro, L. B., Verdu, P., Heyer, E., and Quintana-Murci, L. The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. *Nature Communications*, 5 :3163, 2014. ISSN 2041-1723. doi : 10.1038/ncomms4163.

Pedersen, J. S., Valen, E., Velazquez, A. M. V., Parker, B. J., Rasmussen, M., Lindgreen, S., Lilje, B., Tobin, D. J., Kelly, T. K., Vang, S., Andersson, R., Jones, P. A., Hoover, C. A., Tikhonov, A., Prokhortchouk, E., Rubin, E. M., Sandelin, A., Gilbert, M. T. P., Krogh, A., Willerslev, E., and Orlando, L. Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Research*, 24(3) :454–466, Mar. 2014. ISSN 1088-9051. doi : 10.1101/gr.163592.113. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.163592.113.

Peedicayil, J. The role of DNA methylation in the pathogenesis and treatment of cancer. *Current Clinical Pharmacology*, 7(4) :333–340, Nov. 2012. ISSN 2212-3938.

Peischl, S., Dupanloup, I., Kirkpatrick, M., and Excoffier, L. On the accumulation of deleterious mutations during range expansions. *Molecular Ecology*, 22(24) :

5972–5982, Dec. 2013. ISSN 09621083. doi : 10.1111/mec.12524. URL http://doi.wiley.com/10.1111/mec.12524.

Peng, Y., Shi, H., Qi, X.-b., Xiao, C.-j., Zhong, H., Ma, R.-l. Z., and Su, B. The ADH1b Arg47his polymorphism in East Asian populations and expansion of rice domestication in history. *BMC Evolutionary Biology*, 10(1) :15, 2010. ISSN 1471-2148. doi : 10.1186/1471-2148-10-15. URL http://www.biomedcentral.com/1471-2148/10/15.

Perry, G. H. and Dominy, N. J. Evolution of the human pygmy phenotype. *Trends in Ecology & Evolution*, 24(4) :218–225, Apr. 2009. ISSN 0169-5347. doi : 10.1016/j.tree.2008.11.008.

Perry, G. H., Foll, M., Grenier, J.-C., Patin, E., Nedelec, Y., Pacis, A., Barakatt, M., Gravel, S., Zhou, X., Nsobya, S. L., Excoffier, L., Quintana-Murci, L., Dominy, N. J., and Barreiro, L. B. Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers. *Proceedings of the National Academy of Sciences*, 111(35) :E3596–E3603, Sept. 2014. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.1402875111. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1402875111.

Petronis, A. Epigenetics and twins : three variations on the theme. *Trends in Genetics*, 22(7) :347–350, July 2006. ISSN 01689525. doi : 10.1016/j.tig.2006.04.010. URL http://linkinghub.elsevier.com/retrieve/pii/S0168952506001260.

Pezic, D., Manakov, S. A., Sachidanandam, R., and Aravin, A. A. piRNA pathway targets active LINE1 elements to establish the repressive H3k9me3 mark in germ cells. *Genes & Development*, 28(13) :1410–1428, July 2014. ISSN 0890-9369. doi : 10.1101/gad.240895.114. URL http://genesdev.cshlp.org/cgi/doi/10.1101/gad.240895.114.

Phillips, J. E. and Corces, V. G. CTCF : master weaver of the genome. *Cell*, 137(7) : 1194–1211, June 2009. ISSN 1097-4172. doi : 10.1016/j.cell.2009.06.001.

Picascia, A., Grimaldi, V., Pignalosa, O., De Pascale, M. R., Schiano, C., and Napoli, C. Epigenetic control of autoimmune diseases : From bench to bedside. *Clinical Immunology (Orlando, Fla.)*, 157(1) :1–15, Mar. 2015. ISSN 1521-7035. doi : 10.1016/j.clim.2014.12.013.

Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Srinivasan, B. S., Barsh, G. S., Myers, R. M., Feldman, M. W., and Pritchard, J. K. Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, May 2009. doi : 10.1101/gr.087577.108. URL http://genome.cshlp.org/content/early/2009/03/20/gr.087577.108.abstract.

Pinney, S. Mammalian Non-CpG Methylation : Stem Cells and Beyond. *Biology*, 3 (4) :739–751, Nov. 2014. ISSN 2079-7737. doi : 10.3390/biology3040739. URL http://www.mdpi.com/2079-7737/3/4/739/.

Plagnol, V. and Wall, J. D. Possible Ancestral Structure in Human Populations. *PLoS Genetics*, 2(7) :e105, 2006. ISSN 1553-7390, 1553-7404. doi : 10.1371/journal.pgen.0020105. URL http://dx.plos.org/10.1371/journal.pgen.0020105.

Pond, S. L. K., Frost, S. D. W., and Muse, S. V. HyPhy : hypothesis testing using phylogenies. *Bioinformatics (Oxford, England)*, 21(5) :676–679, Mar. 2005. ISSN 1367-4803. doi : 10.1093/bioinformatics/bti079.

Prezeworski, M., Coop, G., and Wall, J. D. THE SIGNATURE OF POSITIVE SELECTION ON STANDING GENETIC VARIATION. *Evolution*, 59(11) :2312–2323, Nov. 2005. ISSN 0014-3820, 1558-5646. doi : 10.1111/j.0014-3820.2005.tb00941.x. URL http://doi.wiley.com/10.1111/j.0014-3820.2005.tb00941.x.

Pritchard, J. K. and Di Rienzo, A. Adaptation – not by sweeps alone. *Nat Rev Genet*, 11(10) :665–667, Oct. 2010. ISSN 1471-0056. doi : 10.1038/nrg2880. URL http://dx.doi.org/10.1038/nrg2880.

Pritchard, J. K., Pickrell, J. K., and Coop, G. The genetics of human adaptation : hard sweeps, soft sweeps, and polygenic adaptation. *Current biology : CB*, 20(4) : R208–215, Feb. 2010. ISSN 1879-0445. doi : 10.1016/j.cub.2009.11.055.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., Siebauer, M., Theunert, C., Tandon, A., Moorjani, P., Pickrell, J., Mullikin, J. C., Vohr, S. H., Green, R. E., Hellmann, I., Johnson, P. L. F., Blanche, H., Cann, H., Kitzman, J. O., Shendure, J., Eichler, E. E., Lein, E. S., Bakken, T. E., Golovanova, L. V., Doronichev, V. B., Shunkov, M. V., Derevianko, A. P., Viola, B., Slatkin, M., Reich, D., Kelso, J., and Pääbo, S. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481) :43–49, Jan. 2014. ISSN 1476-4687. doi : 10.1038/nature12886.

Ptak, S. E., Hinds, D. A., Koehler, K., Nickel, B., Patil, N., Ballinger, D. G., Przeworski, M., Frazer, K. A., and Pääbo, S. Fine-scale recombination patterns differ between chimpanzees and humans. *Nature Genetics*, 37(4) :429–434, Apr. 2005. ISSN 1061-4036. doi : 10.1038/ng1529. URL http://www.nature.com/doifinder/10.1038/ng1529.

Quintana-Murci, L. and Clark, A. G. Population genetic tools for dissecting innate immunity in humans. *Nature Reviews Immunology*, 13(4) :280–293, Mar. 2013. ISSN 1474-1733, 1474-1741. doi : 10.1038/nri3421. URL http://www.nature.com/doifinder/10.1038/nri3421.

Quintana-Murci, L., Semino, O., Bandelt, H. J., Passarino, G., McElreavey, K., and Santachiara-Benerecetti, A. S. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nature Genetics*, 23(4) :437–441, Dec. 1999. ISSN 1061-4036. doi : 10.1038/70550.

Quintana-Murci, L., Quach, H., Harmant, C., Luca, F., Massonnet, B., Patin, E., Sica, L., Mouguiama-Daouda, P., Comas, D., Tzur, S., Balanovsky, O., Kidd, K. K., Kidd, J. R., van der Veen, L., Hombert, J.-M., Gessain, A., Verdu, P., Froment, A., Bahuchet, S., Heyer, E., Dausset, J., Salas, A., and Behar, D. M. Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proceedings of the National Academy of Sciences of the United States of America*, 105(5) :1596–1601, Feb. 2008. ISSN 1091-6490. doi : 10.1073/pnas.0711467105.

Raj, T., Kuchroo, M., Replogle, J. M., Raychaudhuri, S., Stranger, B. E., and De Jager, P. L. Common risk alleles for inflammatory diseases are targets of recent positive selection. *American Journal of Human Genetics*, 92(4) :517–529, Apr. 2013. ISSN 1537-6605. doi : 10.1016/j.ajhg.2013.03.001.

Rakyan, V. K., Down, T. A., Balding, D. J., and Beck, S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet*, 12(8) :529–541, Aug. 2011. ISSN 1471-0056. doi : 10.1038/nrg3000. URL http://dx.doi.org/10.1038/nrg3000.

Ramos, P. S., Shaftman, S. R., Ward, R. C., and Langefeld, C. D. Genes Associated with SLE Are Targets of Recent Positive Selection. *Autoimmune Diseases*, 2014 : 1–11, 2014. ISSN 2090-0422, 2090-0430. doi : 10.1155/2014/203435. URL http://www.hindawi.com/journals/ad/2014/203435/.

Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A. P., and Jaenisch, R. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of the National Academy of Sciences*, 97 (10) :5237–5242, May 2000. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.97.10.5237. URL http://www.pnas.org/cgi/doi/10.1073/pnas.97.10.5237.

Rands, C. M., Meader, S., Ponting, C. P., and Lunter, G. 8.2% of the Human Genome Is Constrained : Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLoS Genetics*, 10(7) :e1004525, July 2014. ISSN 1553-7404. doi : 10.1371/journal.pgen.1004525. URL http://dx.plos.org/10.1371/journal.pgen.1004525.

Raymond, C. K., Kas, A., Paddock, M., Qiu, R., Zhou, Y., Subramanian, S., Chang, J., Palmieri, A., Haugen, E., Kaul, R., and Olson, M. V. Ancient haplotypes of the HLA Class II region. *Genome Research*, 15(9) :1250–1257, Sept. 2005. ISSN 1088-9051. doi : 10.1101/gr.3554305.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzalez, J. R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., and Hurles, M. E. Global variation in copy number in the human genome. *Nature*,

444(7118) :444–454, Nov. 2006. ISSN 0028-0836. doi : 10.1038/nature05329. URL http://dx.doi.org/10.1038/nature05329.

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L. F., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E. E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M. V., Derevianko, A. P., Hublin, J.-J., Kelso, J., Slatkin, M., and Pääbo, S. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327) :1053–1060, Dec. 2010. ISSN 1476-4687. doi : 10.1038/nature09710.

Reich, D., Patterson, N., Kircher, M., Delfin, F., Nandineni, M. R., Pugach, I., Ko, A. M.-S., Ko, Y.-C., Jinam, T. A., Phipps, M. E., Saitou, N., Wollstein, A., Kayser, M., Pääbo, S., and Stoneking, M. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *American Journal of Human Genetics*, 89(4) :516–528, Oct. 2011. ISSN 1537-6605. doi : 10.1016/j.ajhg.2011.09.005.

Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143) :425–432, May 2007. ISSN 0028-0836, 1476-4687. doi : 10.1038/nature05918. URL http://www.nature.com/doifinder/10.1038/nature05918.

Reik, W., Collick, A., Norris, M. L., Barton, S. C., and Surani, M. A. Genomic imprinting determines methylation of parental alleles in transgenic mice. *Nature*, 328(6127) :248–251, July 1987. ISSN 0028-0836. doi : 10.1038/328248a0.

Richerson, P. J. and Boyd, R. *Not by genes alone : how culture transformed human evolution*. University of Chicago Press, Chicago, 2005. ISBN 0226712842.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. Integrative analysis of 111 reference human epigenomes. *Nature*, 518 (7539) :317–330, Feb. 2015. ISSN 1476-4687. doi : 10.1038/nature14248.

Rook, G. A. W., editor. *The hygiene hypothesis and Darwinian medicine*. Progress in inflammation research. Birkhäuser, Basel ; Boston, 2009. ISBN 9783764389024 3764389028.

Rosiermont, F. L'épigénétique, l'hérédité au-delà de l'ADN, Apr. 2012. URL http://lemonde.fr/sciences/article/2012/04/13/l-epigenetique-une-heredite-sans-adn_1684720_1650684.html.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., and Lander, E. S. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909) :832–837, Oct. 2002. ISSN 0028-0836. doi : 10.1038/nature01140. URL http://dx.doi.org/10.1038/nature01140.

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., and Lander, E. S. Positive Natural Selection in the Human Lineage. *Science*, 312(5780) :1614 –1620, June 2006. doi : 10.1126/science.1124309. URL http://www.sciencemag.org/content/312/5780/1614.abstract.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., and Lander, E. S. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164) :913–918, Oct. 2007. ISSN 0028-0836. doi : 10.1038/nature06250. URL http://dx.doi.org/10.1038/nature06250.

Sams, A. and Hawks, J. Patterns of Population Differentiation and Natural Selection on the Celiac Disease Background Risk Network. *PLoS ONE*, 8(7) :e70564, July 2013. ISSN 1932-6203. doi : 10.1371/journal.pone.0070564. URL http://dx.plos.org/10.1371/journal.pone.0070564.

Sankararaman, S., Patterson, N., Li, H., Pääbo, S., and Reich, D. The Date of Interbreeding between Neandertals and Modern Humans. *PLoS Genetics*, 8(10) : e1002947, Oct. 2012. ISSN 1553-7404. doi : 10.1371/journal.pgen.1002947. URL http://dx.plos.org/10.1371/journal.pgen.1002947.

Santos-Lopes, S. S., Pereira, R. W., Wilson, I. J., and Pena, S. D. A Worldwide Phylogeography for the Human X Chromosome. *PLoS ONE*, 2(6) :e557, June 2007. ISSN 1932-6203. doi : 10.1371/journal.pone.0000557. URL http://dx.plos.org/10.1371/journal.pone.0000557.

Sawyer, S. A. and Hartl, D. L. Population genetics of polymorphism and divergence. *Genetics*, 132(4) :1161–1176, Dec. 1992. ISSN 0016-6731.

Saxonov, S., Berg, P., and Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*, 103(5) :1412–1417, Jan. 2006. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.0510310103. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0510310103.

Scarano, E., Iaccarino, M., Grippo, P., and Parisi, E. The heterogeneity of thymine methyl group origin in DNA pyrimidine isostichs of developing sea urchin embryos. *Proceedings of the National Academy of Sciences of the United States of America*, 57(5) :1394–1400, May 1967. ISSN 0027-8424.

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D. Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*, 15(11) :1576 –1583, Nov. 2005. doi : 10.1101/gr.3709305. URL http://genome.cshlp.org/content/15/11/1576.abstract.

Scheinfeldt, L. B. and Tishkoff, S. A. Recent human adaptation : genomic approaches, interpretation and insights. *Nat Rev Genet*, 14(10) :692–702, Oct. 2013. ISSN 1471-0056. URL http://dx.doi.org/10.1038/nrg3604.

Scheinfeldt, L. B., Soi, S., Thompson, S., Ranciaro, A., Woldemeskel, D., Beggs, W., Lambert, C., Jarvis, J. P., Abate, D., Belay, G., and Tishkoff, S. A. Genetic adaptation to high altitude in the Ethiopian highlands. *Genome Biology*, 13(1) :R1, 2012. ISSN 1465-6914. doi : 10.1186/gb-2012-13-1-r1.

Schermelleh, L., Haemmer, A., Spada, F., Rosing, N., Meilinger, D., Rothbauer, U., Cardoso, M. C., and Leonhardt, H. Dynamics of Dnmt1 interaction with the replication machinery and its role in postreplicative maintenance of DNA methylation. *Nucleic Acids Research*, 35(13) :4301–4312, June 2007. ISSN 0305-1048, 1362-4962. doi : 10.1093/nar/gkm432. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkm432.

Schiffels, S. and Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet*, 46(8) :919–925, Aug. 2014. ISSN 1061-4036. URL http://dx.doi.org/10.1038/ng.3015.

Schneider, E., Pliushch, G., El Hajj, N., Galetzka, D., Puhl, A., Schorsch, M., Frauenknecht, K., Riepert, T., Tresch, A., Muller, A. M., Coerdt, W., Zechner, U., and Haaf, T. Spatial, temporal and interindividual epigenetic variation of functionally important DNA methylation patterns. *Nucleic Acids Research*, 38(12) : 3880–3890, July 2010. ISSN 0305-1048, 1362-4962. doi : 10.1093/nar/gkq126. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkq126.

Schraiber, J. G., Mostovoy, Y., Hsu, T. Y., and Brem, R. B. Inferring Evolutionary Histories of Pathway Regulation from Transcriptional Profiling Data. *PLoS Computational Biology*, 9(10) :e1003255, Oct. 2013. ISSN 1553-7358. doi : 10.1371/journal.pcbi.1003255. URL http://dx.plos.org/10.1371/journal.pcbi.1003255.

Schübeler, D. Function and information content of DNA methylation. *Nature*, 517 (7534) :321–326, Jan. 2015. ISSN 1476-4687. doi : 10.1038/nature14192.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A., and Wigler, M. Large-Scale Copy Number Polymorphism in the Human Genome. *Science*,

305(5683) :525–528, July 2004. doi : 10.1126/science.1098918. URL http://www.sciencemag.org/content/305/5683/525.abstract.

Segurel, L., Thompson, E. E., Flutre, T., Lovstad, J., Venkat, A., Margulis, S. W., Moyse, J., Ross, S., Gamble, K., Sella, G., Ober, C., and Przeworski, M. The ABO blood group is a trans-species polymorphism in primates. *Proceedings of the National Academy of Sciences*, 109(45) :18493–18498, Nov. 2012. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.1210603109. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1210603109.

Seisenberger, S., Peat, J. R., Hore, T. A., Santos, F., Dean, W., and Reik, W. Reprogramming DNA methylation in the mammalian life cycle : building and breaking epigenetic barriers. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 368(1609) :20110330–20110330, Nov. 2012. ISSN 0962-8436, 1471-2970. doi : 10.1098/rstb.2011.0330. URL http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2011.0330.

Selmi, C., Lu, Q., and Humble, M. C. Heritability versus the role of the environment in autoimmunity. *Journal of Autoimmunity*, 39(4) :249–252, Dec. 2012. ISSN 1095-9157. doi : 10.1016/j.jaut.2012.07.011.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. dbSNP : the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1) :308–311, Jan. 2001. ISSN 1362-4962.

Shoemaker, R., Deng, J., Wang, W., and Zhang, K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Research*, 20(7) :883–889, July 2010. ISSN 1088-9051. doi : 10.1101/gr.104695.109. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.104695.109.

Shriver, M. D., Kennedy, G. C., Parra, E. J., Lawson, H. A., Sonpar, V., Huang, J., Akey, J. M., and Jones, K. W. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Human Genomics*, 1(4) :274–286, May 2004. ISSN 1473-9542.

Siddle, K. J. and Quintana-Murci, L. The Red Queen's long race : human adaptation to pathogen pressure. *Current Opinion in Genetics & Development*, 29 :31–38, Dec. 2014. ISSN 1879-0380. doi : 10.1016/j.gde.2014.07.004.

Siegfried, Z. and Simon, I. DNA methylation and gene expression. *Wiley Interdisciplinary Reviews : Systems Biology and Medicine*, 2(3) :362–371, May 2010. ISSN 19395094. doi : 10.1002/wsbm.64. URL http://doi.wiley.com/10.1002/wsbm.64.

Sigurdsson, M. I., Smith, A. V., Bjornsson, H. T., and Jonsson, J. J. The distribution of a germline methylation marker suggests a regional mechanism of LINE-1 silencing by the piRNA-PIWI system. *BMC Genetics*, 13(1) :31, 2012. ISSN 1471-2156. doi : 10.1186/1471-2156-13-31. URL http://www.biomedcentral.com/1471-2156/13/31.

Simons, Y. B., Turchin, M. C., Pritchard, J. K., and Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat Genet*, 46(3) :220–224, Mar. 2014. ISSN 1061-4036. URL http://dx.doi.org/10.1038/ng.2896.

Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., Bai, Z., Lorenzo, F. R., Xing, J., Jorde, L. B., Prchal, J. T., and Ge, R. Genetic evidence for high-altitude adaptation in Tibet. *Science (New York, N.Y.)*, 329(5987) :72–75, July 2010. ISSN 1095-9203. doi : 10.1126/science.1189406.

Sironi, M. and Clerici, M. The hygiene hypothesis : an evolutionary perspective. *Microbes and Infection*, 12(6) :421–427, June 2010. ISSN 12864579. doi : 10.1016/j.micinf.2010.02.002. URL http://linkinghub.elsevier.com/retrieve/pii/S128645791000050X.

Skinner, M. K. Endocrine Disruptors and Epigenetic Transgenerational Disease Etiology. *Pediatric Research*, 61(5 Part 2) :48R–50R, May 2007. ISSN 0031-3998, 1530-0447. doi : 10.1203/pdr.0b013e3180457671. URL http://www.nature.com/doifinder/10.1203/pdr.0b013e3180457671.

Smallwood, S. A. and Kelsey, G. De novo DNA methylation : a germ cell perspective. *Trends in Genetics*, 28(1) :33–42, Jan. 2012. ISSN 01689525. doi : 10.1016/j.tig.2011.09.004. URL http://linkinghub.elsevier.com/retrieve/pii/S0168952511001582.

Smith, A. K., Kilaru, V., Kocak, M., Almli, L. M., Mercer, K. B., Ressler, K. J., Tylavsky, F. A., and Conneely, K. N. Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC genomics*, 15 :145, 2014. ISSN 1471-2164. doi : 10.1186/1471-2164-15-145.

Smith, Z. D. and Meissner, A. DNA methylation : roles in mammalian development. *Nature Reviews Genetics*, 14(3) :204–220, Feb. 2013. ISSN 1471-0056, 1471-0064. doi : 10.1038/nrg3354. URL http://www.nature.com/doifinder/10.1038/nrg3354.

Soranzo, N., Bufe, B., Sabeti, P. C., Wilson, J. F., Weale, M. E., Marguerie, R., Meyerhof, W., and Goldstein, D. B. Positive selection on a high-sensitivity allele of the human bitter-taste receptor TAS2r16. *Current biology : CB*, 15(14) :1257–1265, July 2005. ISSN 0960-9822. doi : 10.1016/j.cub.2005.06.042.

Stankiewicz, P. and Lupski, J. R. Structural Variation in the Human Genome and its Role in Disease. *Annual Review of Medicine*, 61(1) : 437–455, Feb. 2010. ISSN 0066-4219, 1545-326X. doi : 10.1146/annurev-med-100708-204735. URL http://www.annualreviews.org/doi/abs/10.1146/annurev-med-100708-204735.

Strachan, D. P. Family size, infection and atopy : the first decade of the "hygiene hypothesis". *Thorax*, 55 Suppl 1 :S2–10, Aug. 2000. ISSN 0040-6376.

Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Jakobsdottir, M., Steinberg, S., Gudjonsson, S. A., Palsson, A., Thorleifsson, G., Pálsson, S.,

Sigurgeirsson, B., Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K. R., Aben, K. K., Vermeulen, S. H., Goldstein, A. M., Tucker, M. A., Kiemeney, L. A., Olafsson, J. H., Gulcher, J., Kong, A., Thorsteinsdottir, U., and Stefansson, K. Two newly identified genetic determinants of pigmentation in Europeans. *Nature Genetics*, 40(7) :835–837, July 2008. ISSN 1546-1718. doi : 10.1038/ng.160.

Sun, Y. V. The Influences of Genetic and Environmental Factors on Methylome-wide Association Studies for Human Diseases. *Current Genetic Medicine Reports*, 2(4) : 261–270, Dec. 2014. ISSN 2167-4876. doi : 10.1007/s40142-014-0058-2.

Suter, M. A., Anders, A. M., and Aagaard, K. M. Maternal smoking as a model for environmental epigenetic changes affecting birthweight and fetal programming. *Molecular Human Reproduction*, 19(1) :1–6, Jan. 2013. ISSN 1360-9947, 1460-2407. doi : 10.1093/molehr/gas050. URL http://www.molehr.oxfordjournals.org/cgi/doi/10.1093/molehr/gas050.

Szyf, M. The early life social environment and DNA methylation : DNA methylation mediating the long-term impact of social environments early in life. *Epigenetics*, 6(8) :971–978, Aug. 2011. ISSN 1559-2294, 1559-2308. doi : 10.4161/epi.6.8.16793. URL http://www.tandfonline.com/doi/abs/10.4161/epi.6.8.16793.

Ségurel, L. and Quintana-Murci, L. Preserving immune diversity through ancient inheritance and admixture. *Current Opinion in Immunology*, 30 :79–84, Oct. 2014. ISSN 1879-0372. doi : 10.1016/j.coi.2014.08.002.

Ségurel, L., Gao, Z., and Przeworski, M. Ancestry runs deeper than blood : The evolutionary history of *ABO* points to cryptic variation of functional importance : Insights &amp ; Perspective. *BioEssays*, pages n/a–n/a, July 2013. ISSN 02659247. doi : 10.1002/bies.201300030. URL http://doi.wiley.com/10.1002/bies.201300030.

Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L. M., Liu, D. R., Aravind, L., and Rao, A. Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science*, 324(5929) :930–935, May 2009. ISSN 0036-8075, 1095-9203. doi : 10.1126/science.1170116. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1170116.

Tajima, F. EVOLUTIONARY RELATIONSHIP OF DNA SEQUENCES IN FINITE POPULATIONS. *Genetics*, 105(2) :437 –460, Oct. 1983. URL http://www.genetics.org/content/105/2/437.abstract.

Tajima, F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*, 123(3) :585 –595, Nov. 1989. URL http://www.genetics.org/content/123/3/585.abstract.

Tang, J., Xiong, Y., Zhou, H.-H., and Chen, X.-P. DNA methylation and personalized medicine. *Journal of Clinical Pharmacy and Therapeutics*, 39(6) :621–627, Dec. 2014. ISSN 1365-2710. doi : 10.1111/jcpt.12206.

Tang, K., Thornton, K. R., and Stoneking, M. A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. *PLoS Biology*, 5(7) :e171, 2007. ISSN 1544-9173, 1545-7885. doi : 10.1371/journal.pbio.0050171. URL http://biology.plosjournals.org/perlserv/?request=get-document&doi=10.1371%2Fjournal.pbio.0050171.

Tarry-Adkins, J. L. and Ozanne, S. E. The impact of early nutrition on the ageing trajectory. *The Proceedings of the Nutrition Society*, 73(2) :289–301, May 2014. ISSN 1475-2719. doi : 10.1017/S002966511300387X.

Teh, A. L., Pan, H., Chen, L., Ong, M.-L., Dogra, S., Wong, J., MacIsaac, J. L., Mah, S. M., McEwen, L. M., Saw, S.-M., Godfrey, K. M., Chong, Y.-S., Kwek, K., Kwoh, C.-K., Soh, S.-E., Chong, M. F. F., Barton, S., Karnani, N., Cheong, C. Y., Buschdorf, J. P., Stünkel, W., Kobor, M. S., Meaney, M. J., Gluckman, P. D., and Holbrook, J. D. The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. *Genome Research*, 24(7) :1064–1074, July 2014. ISSN 1549-5469. doi : 10.1101/gr.171439.113.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., Sunyaev, S., Bustamante, C. D., Bamshad, M. J., Akey, J. M., Broad GO, Seattle GO, and NHLBI Exome Sequencing Project. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science (New York, N.Y.)*, 337(6090) :64–69, July 2012. ISSN 1095-9203. doi : 10.1126/science.1219240.

Teshima, K. M. and Przeworski, M. Directional positive selection on an allele of arbitrary dominance. *Genetics*, 172(1) :713–718, Jan. 2006. ISSN 0016-6731. doi : 10.1534/genetics.105.044065.

Teshima, K. M., Coop, G., and Przeworski, M. How reliable are empirical genomic scans for selective sweeps ? *Genome Research*, 16(6) :702 –712, June 2006. doi : 10.1101/gr.5105206. URL http://genome.cshlp.org/content/16/6/702.abstract.

The 1000 Genomes Project. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319) :1061–1073, Oct. 2010. ISSN 0028-0836. doi : 10.1038/nature09534. URL http://dx.doi.org/10.1038/nature09534.

The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422) :56–65, Nov. 2012. ISSN 0028-0836. doi : 10.1038/nature11632. URL http://dx.doi.org/10.1038/nature11632.

The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311) :52–58, 2010. ISSN 0028-0836. doi : 10.1038/nature09298. URL http://dx.doi.org/10.1038/nature09298.

The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063) :1299–1320, Oct. 2005. ISSN 0028-0836. doi : 10.1038/nature04226. URL http://dx.doi.org/10.1038/nature04226.

The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164) :851–861, Oct. 2007. ISSN 0028-0836. doi : 10.1038/nature06258. URL http://dx.doi.org/10.1038/nature06258.

Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. J., and Feldman, M. W. Recent common ancestry of human Y chromosomes : Evidence from DNA sequence data. *Proceedings of the National Academy of Sciences*, 97(13) :7360–7365, June 2000. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.97.13.7360. URL http://www.pnas.org/cgi/doi/10.1073/pnas.97.13.7360.

Tishkoff, S. A. Haplotype Diversity and Linkage Disequilibrium at Human G6pd : Recent Origin of Alleles That Confer Malarial Resistance. *Science*, 293(5529) : 455–462, July 2001. ISSN 00368075, 10959203. doi : 10.1126/science.1061573. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1061573.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A., and Deloukas, P. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*, 39(1) :31–40, Jan. 2007. ISSN 1061-4036. doi : 10.1038/ng1946. URL http://www.nature.com/doifinder/10.1038/ng1946.

Tobi, E. W., Lumey, L. H., Talens, R. P., Kremer, D., Putter, H., Stein, A. D., Slagboom, P. E., and Heijmans, B. T. DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Human Molecular Genetics*, 18(21) :4046–4053, Nov. 2009. ISSN 0964-6906, 1460-2083. doi : 10.1093/hmg/ddp353. URL http://www.hmg.oxfordjournals.org/cgi/doi/10.1093/hmg/ddp353.

Toledo-Rodriguez, M., Lotfipour, S., Leonard, G., Perron, M., Richer, L., Veillette, S., Pausova, Z., and Paus, T. Maternal smoking during pregnancy is associated with epigenetic modifications of the brain-derived neurotrophic factor-6 exon in adolescent offspring. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics : The Official Publication of the International Society of Psychiatric Genetics*, 153B(7) :1350–1354, Oct. 2010. ISSN 1552-485X. doi : 10.1002/ajmg.b.31109.

Toperoff, G., Kark, J. D., Aran, D., Nassar, H., Ahmad, W. A., Sinnreich, R., Azaiza, D., Glaser, B., and Hellman, A. Premature aging of leukocyte DNA methylation is associated with type 2 diabetes prevalence. *Clinical Epigenetics*, 7(1) :35, 2015. ISSN 1868-7075. doi : 10.1186/s13148-015-0069-1.

Torgerson, D. G., Boyko, A. R., Hernandez, R. D., Indap, A., Hu, X., White, T. J., Sninsky, J. J., Cargill, M., Adams, M. D., Bustamante, C. D., and Clark, A. G. Evolutionary processes acting on candidate cis-regulatory regions in humans

inferred from patterns of polymorphism and divergence. *PLoS genetics*, 5(8) : e1000592, Aug. 2009. ISSN 1553-7404. doi : 10.1371/journal.pgen.1000592.

Traherne, J. A., Horton, R., Roberts, A. N., Miretti, M. M., Hurles, M. E., Stewart, C. A., Ashurst, J. L., Atrazhev, A. M., Coggill, P., Palmer, S., Almeida, J., Sims, S., Wilming, L. G., Rogers, J., de Jong, P. J., Carrington, M., Elliott, J. F., Sawcer, S., Todd, J. A., Trowsdale, J., and Beck, S. Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history. *PLoS genetics*, 2(1) :e9, Jan. 2006. ISSN 1553-7404. doi : 10.1371/journal.pgen.0020009.

Travis, J. M. J., Münkemüller, T., Burton, O. J., Best, A., Dytham, C., and Johst, K. Deleterious mutations can surf to high densities on the wave front of an expanding population. *Molecular Biology and Evolution*, 24(10) :2334–2343, Oct. 2007. ISSN 0737-4038. doi : 10.1093/molbev/msm167.

Tsaprouni, L. G., Yang, T.-P., Bell, J., Dick, K. J., Kanoni, S., Nisbet, J., Viñuela, A., Grundberg, E., Nelson, C. P., Meduri, E., Buil, A., Cambien, F., Hengstenberg, C., Erdmann, J., Schunkert, H., Goodall, A. H., Ouwehand, W. H., Dermitzakis, E., Spector, T. D., Samani, N. J., and Deloukas, P. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics : official journal of the DNA Methylation Society*, 9(10) : 1382–1396, 2014. ISSN 1559-2308. doi : 10.4161/15592294.2014.969637.

Turchin, M. C., Chiang, C. W., Palmer, C. D., Sankararaman, S., Reich, D., and Hirschhorn, J. N. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat Genet*, 44(9) :1015–1019, Sept. 2012. ISSN 1061-4036. doi : 10.1038/ng.2368. URL http://dx.doi.org/10.1038/ng.2368.

Tysk, C., Lindberg, E., Jarnerot, G., and Floderus-Myrhed, B. Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. *Gut*, 29(7) :990–996, July 1988. ISSN 0017-5749. doi : 10.1136/gut.29.7.990. URL http://gut.bmj.com/cgi/doi/10.1136/gut.29.7.990.

Valinluck, V. and Sowers, L. C. Endogenous Cytosine Damage Products Alter the Site Selectivity of Human DNA Maintenance Methyltransferase DNMT1. *Cancer Research*, 67(3) :946–950, Feb. 2007. ISSN 0008-5472, 1538-7445. doi : 10.1158/0008-5472.CAN-06-3123. URL http://cancerres.aacrjournals.org/cgi/doi/10.1158/0008-5472.CAN-06-3123.

Vallender, E. J. Positive selection on the human genome. *Human Molecular Genetics*, 13(suppl_2) :R245–R254, Oct. 2004. ISSN 1460-2083. doi : 10.1093/hmg/ddh253. URL http://www.hmg.oupjournals.org/cgi/doi/10.1093/hmg/ddh253.

Vandiver, A. R., Irizarry, R. A., Hansen, K. D., Garza, L. A., Runarsson, A., Li, X., Chien, A. L., Wang, T. S., Leung, S. G., Kang, S., and Feinberg, A. P. Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin. *Genome Biology*, 16(1) :80, Apr. 2015. ISSN 1465-6914. doi : 10.1186/s13059-015-0644-y.

Varki, A. and Altheide, T. K. Comparing the human and chimpanzee genomes : Searching for needles in a haystack. *Genome Research*, 15(12) :1746–1758, Dec. 2005. doi : 10.1101/gr.3737405. URL http://genome.cshlp.org/content/15/12/1746.abstract.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507) :

1304–1351, Feb. 2001. ISSN 0036-8075. doi : 10.1126/science.1058040.

Verdu, P., Austerlitz, F., Estoup, A., Vitalis, R., Georges, M., Théry, S., Froment, A., Le Bomin, S., Gessain, A., Hombert, J.-M., Van der Veen, L., Quintana-Murci, L., Bahuchet, S., and Heyer, E. Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Current biology : CB*, 19(4) :312–318, Feb. 2009. ISSN 1879-0445. doi : 10.1016/j.cub.2008.12.049.

Verdu, P., Becker, N. S. A., Froment, A., Georges, M., Grugni, V., Quintana-Murci, L., Hombert, J.-M., Van der Veen, L., Le Bomin, S., Bahuchet, S., Heyer, E., and Austerlitz, F. Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies. *Molecular Biology and Evolution*, 30(4) :918–937, Apr. 2013. ISSN 1537-1719. doi : 10.1093/molbev/mss328.

Vergnaud, G. and Denoeud, F. Minisatellites : Mutability and Genome Architecture. *Genome Research*, 10(7) :899–907, July 2000. doi : 10.1101/gr.10.7.899. URL http://genome.cshlp.org/content/10/7/899.abstract.

Vernot, B. and Akey, J. Complex History of Admixture between Modern Humans and Neandertals. *The American Journal of Human Genetics*, 96(3) :448–453, Mar. 2015. ISSN 00029297. doi : 10.1016/j.ajhg.2015.01.006. URL http://linkinghub.elsevier.com/retrieve/pii/S0002929715000142.

Vernot, B. and Akey, J. M. Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. *Science*, 343(6174) :1017–1021, Feb. 2014. ISSN 0036-8075, 1095-9203. doi : 10.1126/science.1245938. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1245938.

Vernot, B., Stergachis, A. B., Maurano, M. T., Vierstra, J., Neph, S., Thurman, R. E., Stamatoyannopoulos, J. A., and Akey, J. M. Personal and population genomics of human regulatory variation. *Genome Research*, 22(9) :1689–1697, Sept. 2012. ISSN 1549-5469. doi : 10.1101/gr.134890.111.

Vitti, J. J., Grossman, S. R., and Sabeti, P. C. Detecting natural selection in genomic data. *Annual Review of Genetics*, 47 :97–120, 2013. ISSN 1545-2948. doi : 10.1146/annurev-genet-111212-133526.

Voight, B. F., Adams, A. M., Frisse, L. A., Qian, Y., Hudson, R. R., and Di Rienzo, A. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51) :18508–18513, Dec. 2005. doi : 10.1073/pnas.0507325102. URL http://www.pnas.org/content/102/51/18508.abstract.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. A Map of Recent Positive Selection in the Human Genome. *PLoS Biol*, 4(3) :e72, 2006. doi : 10.1371/journal.pbio.0040072. URL http://dx.doi.org/10.1371%2Fjournal.pbio.0040072.

Waddington, C. H. Canalization of Development and the Inheritance of Acquired Characters. *Nature*, 150(3811) :563–565, Nov. 1942. ISSN 0028-0836. doi : 10.1038/150563a0. URL http://www.nature.com/doifinder/10.1038/150563a0.

Wagner, J. R., Busche, S., Ge, B., Kwan, T., Pastinen, T., and Blanchette, M. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biology*, 15(2) :R37, 2014. ISSN 1465-6906. doi : 10.1186/gb-2014-15-2-r37. URL http://genomebiology.com/2014/15/2/R37.

Wall, J. D., Yang, M. A., Jay, F., Kim, S. K., Durand, E. Y., Stevison, L. S., Gignoux, C., Woerner, A., Hammer, M. F., and Slatkin, M. Higher Levels of Neanderthal Ancestry in East Asians than in Europeans. *Genetics*, 194(1) :199–209, May 2013. ISSN 0016-6731. doi : 10.1534/genetics.112.148213. URL http://www.genetics.org/cgi/doi/10.1534/genetics.112.148213.

Wang, E. T., Kodama, G., Baldi, P., and Moyzis, R. K. Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proceedings of the National Academy of Sciences*, 103(1) :135–140, Jan. 2006. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.0509691102. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0509691102.

Ward, L. D. and Kellis, M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science (New York, N.Y.)*, 337(6102) :1675–1678, Sept. 2012a. ISSN 1095-9203. doi : 10.1126/science.1225057.

Ward, L. D. and Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotech*, 30(11) :1095–1106, Nov. 2012b. ISSN 1087-0156. URL http://dx.doi.org/10.1038/nbt.2422.

Watanabe, T., Tomizawa, S.-i., Mitsuya, K., Totoki, Y., Yamamoto, Y., Kuramochi-Miyagawa, S., Iida, N., Hoki, Y., Murphy, P. J., Toyoda, A., Gotoh, K., Hiura, H., Arima, T., Fujiyama, A., Sado, T., Shibata, T., Nakano, T., Lin, H., Ichiyanagi, K., Soloway, P. D., and Sasaki, H. Role for piRNAs and noncoding RNA in de novo DNA methylation of the imprinted mouse Rasgrf1 locus. *Science (New York, N.Y.)*, 332(6031) :848–852, May 2011. ISSN 1095-9203. doi : 10.1126/science.1203919.

Weedon, M. N., Lettre, G., Freathy, R. M., Lindgren, C. M., Voight, B. F., Perry, J. R. B., Elliott, K. S., Hackett, R., Guiducci, C., Shields, B., Zeggini, E., Lango, H., Lyssenko, V., Timpson, N. J., Burtt, N. P., Rayner, N. W., Saxena, R., Ardlie, K., Tobias, J. H., Ness, A. R., Ring, S. M., Palmer, C. N. A., Morris, A. D., Peltonen, L., Salomaa, V., Diabetes Genetics Initiative, Wellcome Trust Case Control Consortium, Davey Smith, G., Groop, L. C., Hattersley, A. T., McCarthy, M. I., Hirschhorn, J. N., and Frayling, T. M. A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nature Genetics*, 39(10) :1245–1250, Oct. 2007. ISSN 1546-1718. doi : 10.1038/ng2121.

Weidner, C., Lin, Q., Koch, C., Eisele, L., Beier, F., Ziegler, P., Bauerschlag, D., Jöckel, K.-H., Erbel, R., Mühleisen, T., Zenke, M., Brümmendorf, T., and Wagner,

W. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biology*, 15(2) :R24, 2014. ISSN 1465-6906. doi : 10.1186/gb-2014-15-2-r24. URL http://genomebiology.com/2014/15/2/R24.

Weir, B. S. and Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38(6) :1358, Nov. 1984. ISSN 00143820. doi : 10.2307/2408641. URL http://www.jstor.org/stable/2408641?origin=crossref.

Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M., and Hill, W. G. Measures of human population structure show heterogeneity among genomic regions. *Genome Research*, 15(11) :1468–1476, Nov. 2005. ISSN 1088-9051. doi : 10.1101/gr.4398405.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1) :D1001–D1006, Jan. 2014. doi : 10.1093/nar/gkt1229. URL http://nar.oxfordjournals.org/content/42/D1/D1001.abstract.

Wilde, S., Timpson, A., Kirsanow, K., Kaiser, E., Kayser, M., Unterländer, M., Hollfelder, N., Potekhina, I. D., Schier, W., Thomas, M. G., and Burger, J. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proceedings of the National Academy of Sciences*, 111 (13) :4832–4837, Apr. 2014. ISSN 0027-8424, 1091-6490. doi : 10.1073/pnas.1316513111. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1316513111.

Willett, W. C. Balancing life-style and genomics research for disease prevention. *Science (New York, N.Y.)*, 296(5568) :695–698, Apr. 2002. ISSN 1095-9203. doi : 10.1126/science.1071055.

Williamson, S. H., Hubisz, M. J., Clark, A. G., Payseur, B. A., Bustamante, C. D., and Nielsen, R. Localizing recent adaptive evolution in the human genome. *PLoS genetics*, 3(6) :e90, June 2007. ISSN 1553-7404. doi : 10.1371/journal.pgen.0030090.

Winckler, W. Comparison of Fine-Scale Recombination Rates in Humans and Chimpanzees. *Science*, 308(5718) :107–111, Apr. 2005. ISSN 0036-8075, 1095-9203. doi : 10.1126/science.1105322. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1105322.

Wlasiuk, G., Khan, S., Switzer, W. M., and Nachman, M. W. A history of recurrent positive selection at the toll-like receptor 5 in primates. *Molecular Biology and Evolution*, 26(4) :937–949, Apr. 2009. ISSN 1537-1719. doi : 10.1093/molbev/msp018.

Wolffe, A. P. and Guschin, D. Review : Chromatin Structural Features and Targets That Regulate Transcription. *Journal of Structural Biology*, 129(2-3) :102–122, Apr. 2000. ISSN 10478477. doi : 10.1006/jsbi.2000.4217. URL http://linkinghub.elsevier.com/retrieve/pii/S1047847700942175.

Wollstein, A. and Stephan, W. Inferring positive selection in humans from genomic data. *Investigative Genetics*, 6(1), Dec. 2015. ISSN 2041-2223. doi : 10.1186/s13323-015-0023-1. URL http://www.investigativegenetics.com/content/6/1/5.

Wong, A. H. Phenotypic differences in genetically identical organisms : the epigenetic perspective. *Human Molecular Genetics*, 14(suppl_1) :R11–R18, Apr. 2005. ISSN 0964-6906, 1460-2083. doi : 10.1093/hmg/ddi116. URL http://www.hmg.oupjournals.org/cgi/doi/10.1093/hmg/ddi116.

Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nature Reviews. Genetics*, 8(3) :206–216, Mar. 2007. ISSN 1471-0056. doi : 10.1038/nrg2063.

Wright, S. Evolution in Mendelian Populations. *Genetics*, 16(2) :97–159, Mar. 1931. ISSN 0016-6731.

Wright, S. Isolation by Distance. *Genetics*, 28(2) :114–138, Mar. 1943. ISSN 0016-6731.

Wright, S. The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. *Evolution*, 19(3) :395, Sept. 1965. ISSN 00143820. doi : 10.2307/2406450. URL http://www.jstor.org/stable/2406450?origin=crossref.

Xu, Q., Xing, S., Zhu, C., Liu, W., Fan, Y., Wang, Q., Song, Z., Yang, W., Luo, F., Shang, F., Kang, L., Chen, W., Yan, J., Li, J., and Sang, T. Population transcriptomics reveals a potentially positive role of expression diversity in adaptation : Expression diversity in adaptation. *Journal of Integrative Plant Biology*, 57(3) :284–299, Mar. 2015. ISSN 16729072. doi : 10.1111/jipb.12287. URL http://doi.wiley.com/10.1111/jipb.12287.

Yan, H., Zhang, D., Liu, H., Wei, Y., Lv, J., Wang, F., Zhang, C., Wu, Q., Su, J., and Zhang, Y. Chromatin modifications and genomic contexts linked to dynamic DNA methylation patterns across human cell types. *Scientific Reports*, 5 :8410, 2015. ISSN 2045-2322. doi : 10.1038/srep08410.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, 42(7) :565–569, July 2010. ISSN 1061-4036. doi : 10.1038/ng.608. URL http://dx.doi.org/10.1038/ng.608.

Yang, n. and Bielawski, n. Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, 15(12) :496–503, Dec. 2000. ISSN 1872-8383.

Yang, Z. PAML 4 : phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8) :1586–1591, Aug. 2007. ISSN 0737-4038. doi : 10.1093/molbev/msm088.

Ye, K., Lu, J., Raj, S. M., and Gu, Z. Human Expression QTLs Are Enriched in Signals of Environmental Adaptation. *Genome Biology and Evolution*, 5(9) :1689–1701, Sept. 2013. ISSN 1759-6653. doi : 10.1093/gbe/evt124. URL http://gbe.oxfordjournals.org/cgi/doi/10.1093/gbe/evt124.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T. S., Zheng, H., Liu, T., He, W., Li, K., Luo, R., Nie, X., Wu, H., Zhao, M., Cao, H., Zou, J., Shan, Y., Li, S., Yang, Q., Asan, Ni, P., Tian, G., Xu, J., Liu, X., Jiang, T., Wu, R., Zhou, G., Tang, M., Qin, J., Wang, T., Feng, S., Li, G., Huasang, Luosang, J., Wang, W., Chen, F., Wang, Y., Zheng, X., Li, Z., Bianba, Z., Yang, G., Wang, X., Tang, S., Gao, G., Chen, Y., Luo, Z., Gusang, L., Cao, Z., Zhang, Q., Ouyang, W., Ren, X., Liang, H., Zheng, H., Huang, Y., Li, J., Bolund, L., Kristiansen, K., Li, Y., Zhang, Y., Zhang, X., Li, R., Li, S., Yang, H., Nielsen, R., Wang, J., and Wang, J. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science*, 329(5987) :75–78, July 2010. doi : 10.1126/science.1190371. URL http://www.sciencemag.org/content/329/5987/75.abstract.

Yu, M., Hon, G., Szulwach, K., Song, C.-X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B., Min, J.-H., Jin, P., Ren, B., and He, C. Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome. *Cell*, 149(6) :1368–1380, June 2012. ISSN 00928674. doi : 10.1016/j.cell.2012.04.027. URL http://linkinghub.elsevier.com/retrieve/pii/S009286741200534X.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., Shen, Y., Pervouchine, D. D., Djebali, S., Thurman, R. E., Kaul, R., Rynes, E., Kirilusha, A., Marinov, G. K., Williams, B. A., Trout, D., Amrhein, H., Fisher-Aylor, K., Antoshechkin, I., DeSalvo, G., See, L.-H., Fastuca, M., Drenkow, J., Zaleski, C., Dobin, A., Prieto, P., Lagarde, J., Bussotti, G., Tanzer, A., Denas, O., Li, K., Bender, M. A., Zhang, M., Byron, R., Groudine, M. T., McCleary, D., Pham, L., Ye, Z., Kuan, S., Edsall, L., Wu, Y.-C., Rasmussen, M. D., Bansal, M. S., Kellis, M., Keller, C. A., Morrissey, C. S., Mishra, T., Jain, D., Dogan, N., Harris, R. S., Cayting, P., Kawli, T., Boyle, A. P., Euskirchen, G., Kundaje, A., Lin, S., Lin, Y., Jansen, C., Malladi, V. S., Cline, M. S., Erickson, D. T., Kirkup, V. M., Learned, K., Sloan, C. A., Rosenbloom, K. R., Lacerda de Sousa, B., Beal, K., Pignatelli, M., Flicek, P., Lian, J., Kahveci, T., Lee, D., James Kent, W., Ramalho Santos, M., Herrero, J., Notredame, C., Johnson, A., Vong, S., Lee, K., Bates, D., Neri, F., Diegel, M., Canfield, T., Sabo, P. J., Wilken, M. S., Reh, T. A., Giste, E., Shafer, A., Kutyavin, T., Haugen, E., Dunn, D., Reynolds, A. P., Neph, S., Humbert, R., Scott Hansen, R., De Bruijn, M., Selleri, L., Rudensky, A., Josefowicz, S., Samstein, R., Eichler, E. E., Orkin, S. H., Levasseur, D., Papayannopoulou, T., Chang, K.-H., Skoultchi, A., Gosh, S., Disteche, C., Treuting, P., Wang, Y., Weiss, M. J., Blobel, G. A., Cao, X., Zhong, S., Wang, T., Good, P. J., Lowdon, R. F., Adams, L. B., Zhou, X.-Q., Pazin, M. J., Feingold, E. A., Wold, B., Taylor, J., Mortazavi, A., Weissman, S. M., Stamatoyannopoulos, J. A., Snyder, M. P., Guigo, R., Gingeras, T. R., Gilbert, D. M., Hardison, R. C., Beer, M. A., Ren, B., and The Mouse ENCODE Consortium. A comparative encyclopedia of DNA elements in

the mouse genome. *Nature*, 515(7527) :355–364, Nov. 2014. ISSN 0028-0836. URL http://dx.doi.org/10.1038/nature13992.

Zaidi, S. K., Young, D. W., Montecino, M., Lian, J. B., Stein, J. L., van Wijnen, A. J., and Stein, G. S. Architectural Epigenetics : Mitotic Retention of Mammalian Transcriptional Regulatory Information. *Molecular and Cellular Biology*, 30(20) : 4758–4766, Oct. 2010. ISSN 0270-7306. doi : 10.1128/MCB.00646-10. URL http://mcb.asm.org/cgi/doi/10.1128/MCB.00646-10.

Zdravkovic, S., Wienke, A., Pedersen, N. L., Marenberg, M. E., Yashin, A. I., and De Faire, U. Heritability of death from coronary heart disease : a 36-year follow-up of 20 966 Swedish twins. *Journal of Internal Medicine*, 252(3) :247–254, Sept. 2002. ISSN 0954-6820, 1365-2796. doi : 10.1046/j.1365-2796.2002.01029.x. URL http://doi.wiley.com/10.1046/j.1365-2796.2002.01029.x.

Zeng, K., Fu, Y.-X., Shi, S., and Wu, C.-I. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, 174(3) :1431–1439, Nov. 2006. ISSN 0016-6731. doi : 10.1534/genetics.106.061432.

Zhang, C., Bailey, D. K., Awad, T., Liu, G., Xing, G., Cao, M., Valmeekam, V., Retief, J., Matsuzaki, H., Taub, M., Seielstad, M., and Kennedy, G. C. A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations. *Bioinformatics (Oxford, England)*, 22(17) :2122–2128, Sept. 2006. ISSN 1367-4811. doi : 10.1093/bioinformatics/btl365.

Zhang, D., Cheng, L., Badner, J. A., Chen, C., Chen, Q., Luo, W., Craig, D. W., Redman, M., Gershon, E. S., and Liu, C. Genetic Control of Individual Differences in Gene-Specific Methylation in Human Brain. *The American Journal of Human Genetics*, 86(3) :411–419, Mar. 2010. ISSN 00029297. doi : 10.1016/j.ajhg.2010.02.005. URL http://linkinghub.elsevier.com/retrieve/pii/S000292971000087X.

Zhang, G., Muglia, L. J., Chakraborty, R., Akey, J. M., and Williams, S. M. Signatures of natural selection on genetic variants affecting complex human traits. *Applied & Translational Genomics*, 2 :78–94, Dec. 2013. ISSN 22120661. doi : 10.1016/j.atg.2013.10.002. URL http://linkinghub.elsevier.com/retrieve/pii/S2212066113000173.

Zhang, J., Nielsen, R., and Yang, Z. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Molecular Biology and Evolution*, 22(12) :2472–2479, Dec. 2005. doi : 10.1093/molbev/msi237. URL http://mbe.oxfordjournals.org/content/22/12/2472.abstract.

Zhang, L., Lu, X., Lu, J., Liang, H., Dai, Q., Xu, G.-L., Luo, C., Jiang, H., and He, C. Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. *Nature Chemical Biology*, 8(4) :328–330, Apr. 2012. ISSN 1552-4469. doi : 10.1038/nchembio.914.

Zhao, B., Yang, Y., Wang, X., Chong, Z., Yin, R., Song, S.-H., Zhao, C., Li, C., Huang, H., Sun, B.-F., Wu, D., Jin, K.-X., Song, M., Zhu, B.-Z., Jiang, G., Rendtlew Danielsen, J. M., Xu, G.-L., Yang, Y.-G., and Wang, H. Redox-active quinones induces genome-wide DNA methylation changes by an iron-mediated and Tet-dependent mechanism. *Nucleic Acids Research*, 42(3) :1593–1605, Feb. 2014. ISSN 1362-4962. doi : 10.1093/nar/gkt1090.

Zhi, D., Aslibekyan, S., Irvin, M. R., Claas, S. A., Borecki, I. B., Ordovas, J. M., Absher, D. M., and Arnett, D. K. SNPs located at CpG sites modulate genome-epigenome interaction. *Epigenetics : official journal of the DNA Methylation Society*, 8(8) :802–806, Aug. 2013. ISSN 1559-2308. doi : 10.4161/epi.25501.

ANNEXES

Annexe A

Compléments d'informations pour l'article 1

Exploring the Occurrence of Classic Selective Sweeps in Humans Using Whole-genome Sequencing Datasets

Supplementary material

Maud Fagny[1,2,3], Etienne Patin[1,2], David Enard[4], Luis B. Barreiro[5], Lluis Quintana-Murci[1,2,6] and Guillaume Laval[1,2,6]

[1] Institut Pasteur, Human Evolutionary Genetics, Department of Genomes and Genetics, F-75015 Paris, France
[2] Centre National de la Recherche Scientifique, URA3012, F-75015 Paris, France
[3] Univ. Pierre et Marie Curie, Cellule Pasteur UPMC, F-75015 Paris, France
[4] Department of Biology, Stanford University, Stanford, CA 94305-5020, USA
[5] Sainte-Justine Hospital Research Center, Department of Pediatrics, University of Montreal, Montreal, Quebec H3T 1C5, Canada

[6] Corresponding authors. E-mail addresses: glaval@pasteur.fr, quintana@pasteur.fr

This file includes:
Supplementary Text
Supplementary Figures S1 to S11
Supplementary Tables S1 to S9, S13 to S15 and S17 to S20.
References

Additional files provided as separate files.
Supplementary Tables S10 to S12 and S16 are provided as separate Excel files, each with multiple worksheets.

Supplementary text

Calibrating the demographic model. We calibrated a model of demography that can be easily simulated using SFS_CODE, to estimate the power to detect selection considering realistic demographic scenarios generally observed in humans, i.e. expansion in Africa and bottleneck/expansion in non-African populations (Laval and Excoffier 2004; Voight et al. 2005; Gravel et al. 2011). We used an Approximate Bayesian Computation (ABC) framework (Beaumont et al. 2002) to infer an approximated model allowing us to reproduce the patterns of diversity observed in human populations. The data used, originally published in (Laval et al. 2010), consist in 95 sub-Saharan African individuals, 47 European individuals, and 48 East-Asian individuals. The Mozambican population was not considered here because of their extreme signal of population expansion, with respect to all other African populations, can bias our model calibration toward extreme expansions. We used 20 autosomal, independent non-coding regions that met criteria determined by the need for genetic variation evolving under neutrality and therefore influenced by demography alone (Laval et al. 2010). For each region, we computed the Tajima's *D*, the number of segregating sites *S* and the F_{ST}, and we used the means of these statistics over the 20 noncoding regions to combine information from multiple loci. To focus on demography, we calibrated only demographic parameters (expansions and bottlenecks) and set population divergence parameters (split time and migration rate) to values previously inferred for these populations and commonly admitted (Laval et al. 2010; Gravel et al. 2011). The ancestral African population of constant size (N=10,000) split into two populations (African and non-African) ~60,000 years ago, with non-Africans splitting again into two populations (European and Asian) ~20,000 years ago (between continent migration rate 1.3×10^{-5}). We then simulated multiple scenarios of expansions in Africans and bottlenecks/expansions in Europeans and Asians (note that we systematically used the same parameters for the two non-African populations). The list of models and the SFS_CODE command lines are reported in the supplementary table S1, Supplementary Material online. For each model, we performed 5×10^{4} simulations of 20 noncoding sequences and we retained the model that presented the highest posterior probability. To estimate the posterior probability of each model, we merged all simulations and retained the best, i.e. the 1% that provided the simulated averaged summary statistics the closest to the empirical averaged summary statistics. The posterior probability of a given model is simply the proportion of the best simulations obtained for the model. The two models that provided the highest posterior probability were virtually the same. They involved an instantaneous expansion in African and non-African populations (Exp1=50 and Exp2=100, respectively, supplementary table S1, Supplementary Material online) and slightly differed for the bottleneck among non-Africans (B=0.5 for the first model and B=0.6 for the second model). We retained arbitrarily the model with B=0.5 (population size divided by 2) for all subsequent simulations (under the neutral and selection models). We next checked by re-simulation if the retained calibrated model allowed us to obtain simulated summary statistics that matched to empirical summary statistics (supplementary fig. S1, Supplementary Material online). We found that this calibrated model well reproduced the genetic diversity observed in human populations according to their continental origin.

Replication of the HapMap enrichment of genic SNPs among iHS outliers. In contrast with the DIND results, no strongly significant enrichments in genic SNPs were found among the iHS outliers (table 1), while such significant enrichments have

been previously detected in all the populations of the HapMap Phase 1 genotyping dataset (Voight et al. 2006). We hypothesized that this observed lack of enrichment of functional SNPs among iHS outliers could result from a number of factors inherent to the methodology used to compute and test the enrichments. To test this, we downloaded the iHS values of the HapMap Phase 2 dataset (Frazer et al. 2007) from the Haplotter website (http://haplotter.uchicago.edu/) and we applied our procedure to compute and test the ORs. We confirmed the enrichments of genic SNPs among the iHS outliers in all populations with values of OR similar to those previously found without applying any correction for confounder (Voight et al. 2006) (supplementary table S8, Supplementary Material online). This indicates that our methodology does not affect the iHS test specifically. However, it is noteworthy that our resampling method leads to *P*-values generally less significant than previously obtained (Voight et al. 2006). This, together with the fact that the enrichment previously observed in Asia becomes non significant with our resampling method, suggests that our statistical approach to test the effects of recent positive selection is conservative (supplementary table S8, Supplementary Material online).

Enrichment of genic SNPs among outliers is not sensitive to window size. We next hypothesized that the window size could also account for the lack of power of iHS in the context of WGS data. Because the EHH associated with the selected allele can persist over 1 Mb for strong selection coefficients ($2Ns>100$) (Voight et al. 2006), the use of 100 kb windows may have led to an underestimation of the strongest signals of positive selection. To test this, we computed iHS, together with DIND, over windows of 1 Mb. No significant enrichments of genic or non-synonymous SNPs among the iHS outliers were detected in any population (supplementary table S9, Supplementary Material online). These observations, together with high genome-wide correlations observed between values of iHS computed over 100 kb and 1 Mb regions (Pearson correlation coefficient *r* equal to 0.989, 0.980 and 0.958 for the 1000G Pilot, Phase1 and the CG datasets, respectively), suggest that our results are not sensitive to the window size. All these results, combined with our simulations results showing that iHS had sufficient power to detect positive selection on 100 kb windows (figs. 2 and 3), indicate that our statistical approach to test the effects of recent positive selection cannot account for the observed lack of enrichment of functional SNPs among iHS outliers. By contrast, significant enrichments were obtained among DIND outliers. This observation may be surprising in light of the fact that, when computing DIND statistics individually, the selection signal (i.e., decreased diversity around the selected mutation due to the sweep) could be diluted over 1 Mb. When performing power simulations, although the power of DIND was slightly decreased in 1 Mb regions, with respect to 100 kb regions, the power remained satisfactory (supplementary fig. S6, Supplementary Material online).

Command lines used

Simulation under neutrality of the retained demographic scenario: ./sfs_code 3 100 -N 100 -L 1 100000 -t 0.001 -r 0.001 -TS 0.08 0 1 -Td 0.08 P 1 0.5 -Tm 0.08 P 0 1 0.2 -Tm 0.08 P 1 0 0.1 -TS 0.16 1 2 -Tm 0.16 L 0.2 0.2 0.1 0.1 0.1 0.1 -Td 0.16 P 0 50 -Tm 0.16 L 10 10 0.1 0.1 0.1 0.1 -Td 0.188 P 1 100 -Td 0.188 P 2 100 -Tm 0.188 L 10 10 10 10 10 10 100 -n 59 60 60 -TE 0.2 -o out_neutrality

Simulation under positive selection considering the retained demographic scenario: ./sfs_code $Neutral_demography --mutation $t_mut P $pop S 50000 G 100 "$Neutral_demography" is the command line used to simulate the retained demographic scenario, see command line above. "$t_mut" is the time at which a new

advantageous mutation was inserted into the middle of the sequence, at a frequency of 1/2*N*, in a specific population, termed "$pop".

Simulation under background selection considering the retained demographic scenario: ./sfs_code $Neutral_demography -W 1 $Gamma 0 0.2
"$Gamma" is the 2*Ns* used to simulate the 20% of mutations negatively selected

Simulation under positive selection together with background selection considering the retained demographic scenario: ./sfs_code $Neutral_demography --mutation $t_mut P $pop S 50000 G 100 -W 1 $Gamma 0 0.2

Simulation under positive selection on standing variation, the mpop command line used is: ./mpop -N $effectivesize -g 100 -O $intialfreq -m 0.001 -r 0.001 -s $selectivecoeff -h 0.5 -i $outfile -e $seed

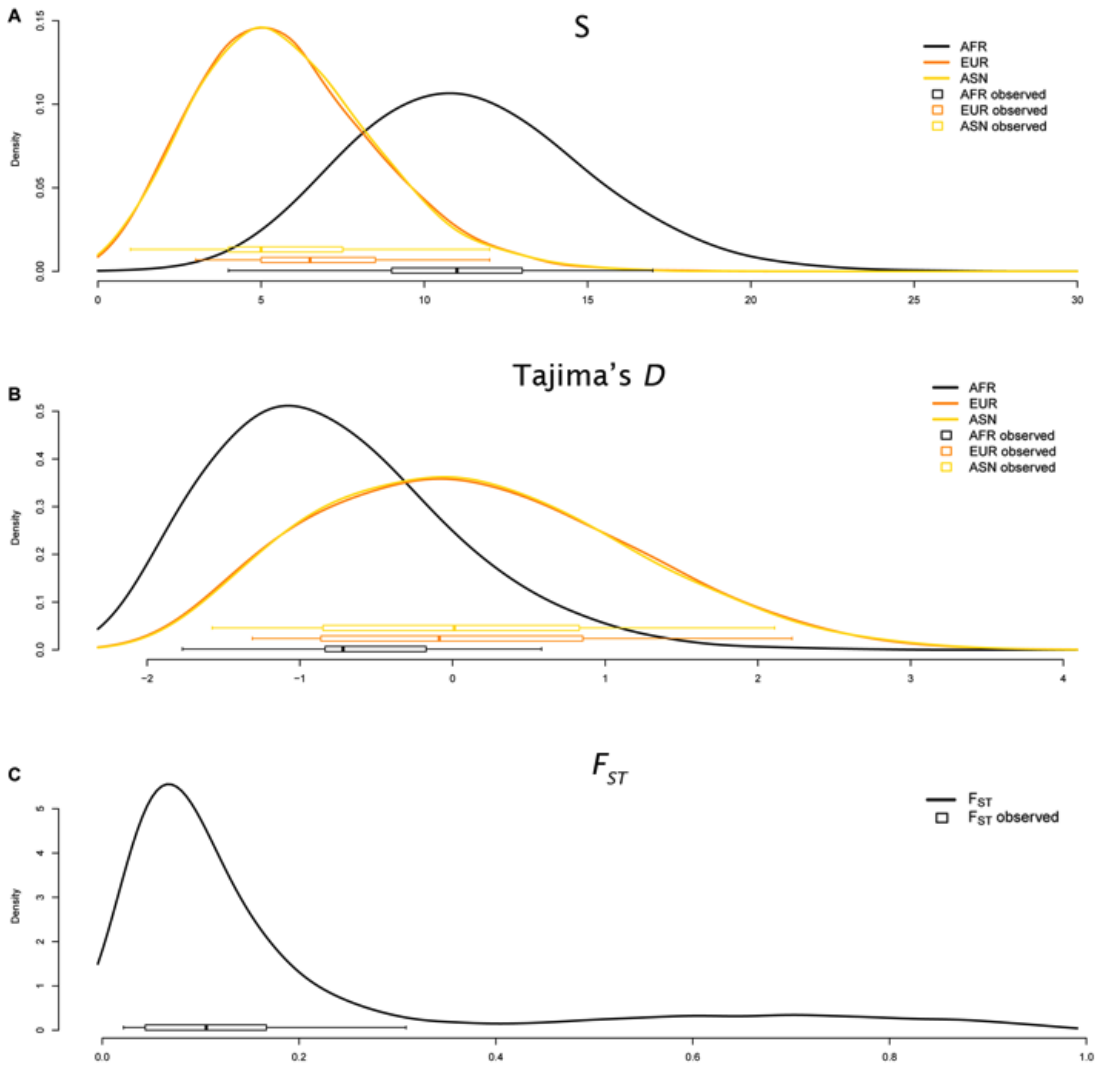


FIG. S1. Distributions of simulated summary statistics in comparison with empirical summary statistics. We performed a total of 2×10^4 simulations (regions of 1.3 Kb) of the calibrated demographic model (supplementary table S1, Supplementary Material online). For each simulation, we computed (*A*) the number of segregating sites *S*, (*B*) Tajima's *D*, and (*C*) F_{ST}, and compared these simulated summary statistics to the empirical summary statistics. The box plots of the empirical summary statistics computed for each of the 20 autosomal, independent non-coding regions are shown.

SFS_CODE command line used to simulate the retained demographic scenario (see Materials and Methods): ./sfs_code 3 100 -N 100 -L 1 1300 -t 0.001 -r 0.001 -TS 0.08 0 1 -Td 0.08 P 1 $B -Tm 0.08 P 0 1 0.2 -Tm 0.08 P 1 0 0.1 -TS 0.16 1 2 -Tm 0.16 L 0.2 0.2 0.1 0.1 0.1 0.1 -Td $t_Exp1 P 0 $Exp1 -Tm 0.16 L 10 10 0.1 0.1 0.1 0.1 -Td 0.188 P 1 $Exp2 -Td 0.188 P 2 $Exp2 -Tm 0.188 L 10 10 10 10 10 10 100 -n 95 60 60 -TE 0.2
The values used for the retained demographic scenario are $B=0.5, $t_Exp1=0.6, $Exp1=50 and $Exp2=100

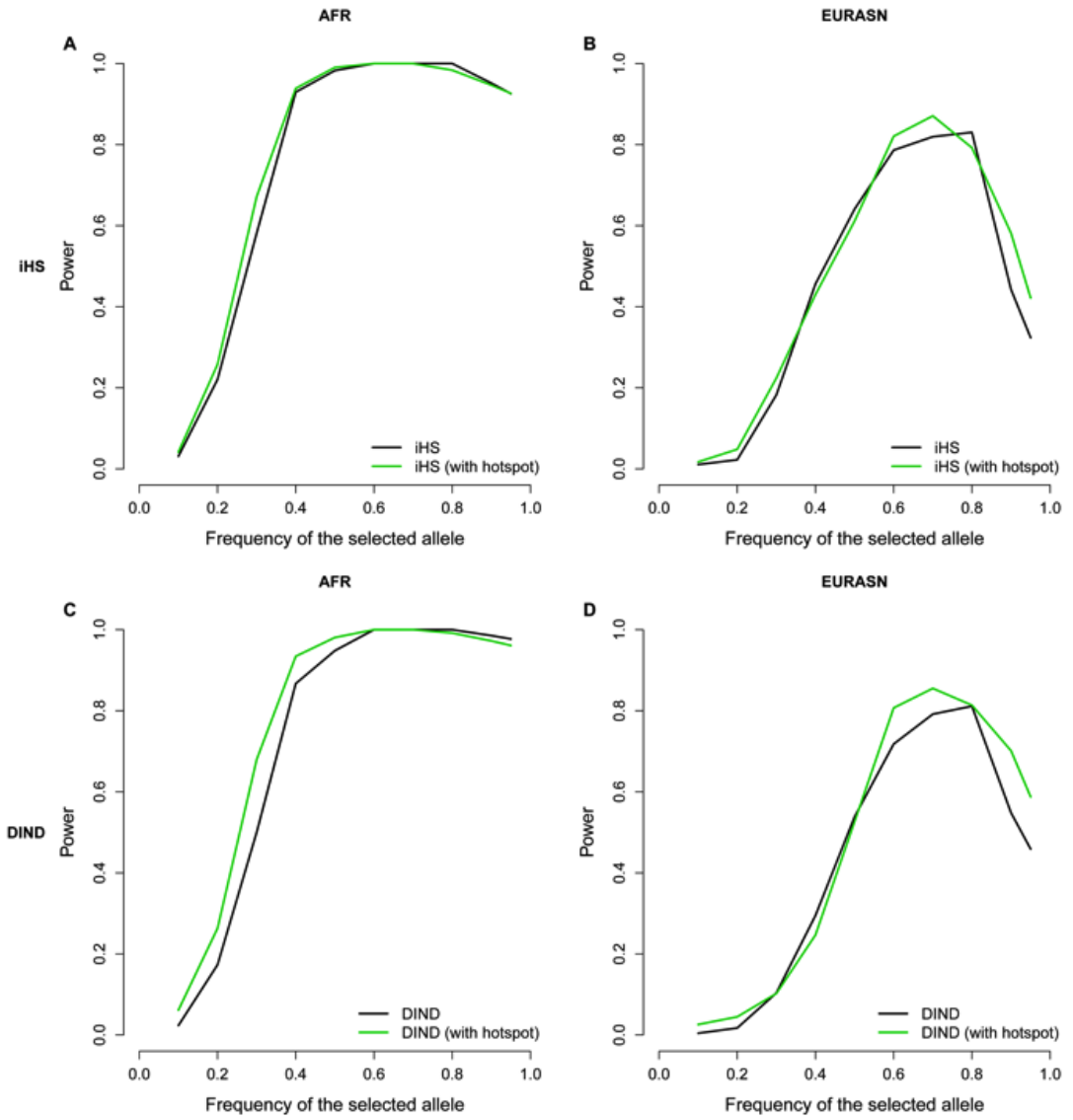


FIG. S2. Power of iHS and DIND to detect hard sweeps in regions where variation of recombination rate was simulated. We simulated "full sequence data" (100 kb) with a SNP under recent positive selection (2*Ns*=100, see Materials and Methods). We simulated 1 recombination hotspot randomly inserted in each region (hotspots occurred at an average rate of 1 per 100 kb in humans), with intensity set to 30-fold the background recombination rate. Black curves correspond to the results shown in figure 2. Green curves correspond to the power of iHS (computed using the simulated genetics maps) and DIND obtained using critical values (FPR=0.01) computed from 10^3 neutral simulations modelling similar hotspots. Power of iHS (*A*-*B*) and DIND (*C*-*D*) computed using the proportion of extreme values (see Materials and Methods). In each case, we performed a total of ~2,000 simulations. (*A*, *C*) African population. (*B*, *D*) Eurasian populations (EURASN): European and Asian populations for which we used the same demographic model.

SFS_CODE command line: ./sfs_code 3 100 -N 100 -L 1 100000 -t 0.001 -r F "maps_recombination.txt" 0.001 -TS 0.08 0 1 -Td 0.08 P 1 0.5 -Tm 0.08 P 0 1 0.2 -Tm 0.08 P 1 0 0.1 -TS 0.16 1 2 -Tm 0.16 L 0.2 0.2 0.1 0.1 0.1 0.1 -Td 0.16 P 0 50 -Tm 0.16 L 10 10 0.1 0.1 0.1 0.1 -Td 0.188 P 1 100 -Td 0.188 P 2 100 -Tm 0.188 L 10 10 10 10 10 10 100 --mutation $t_mut P $pop S 50000 G 100 -n 59 60 60 -TE 0.2

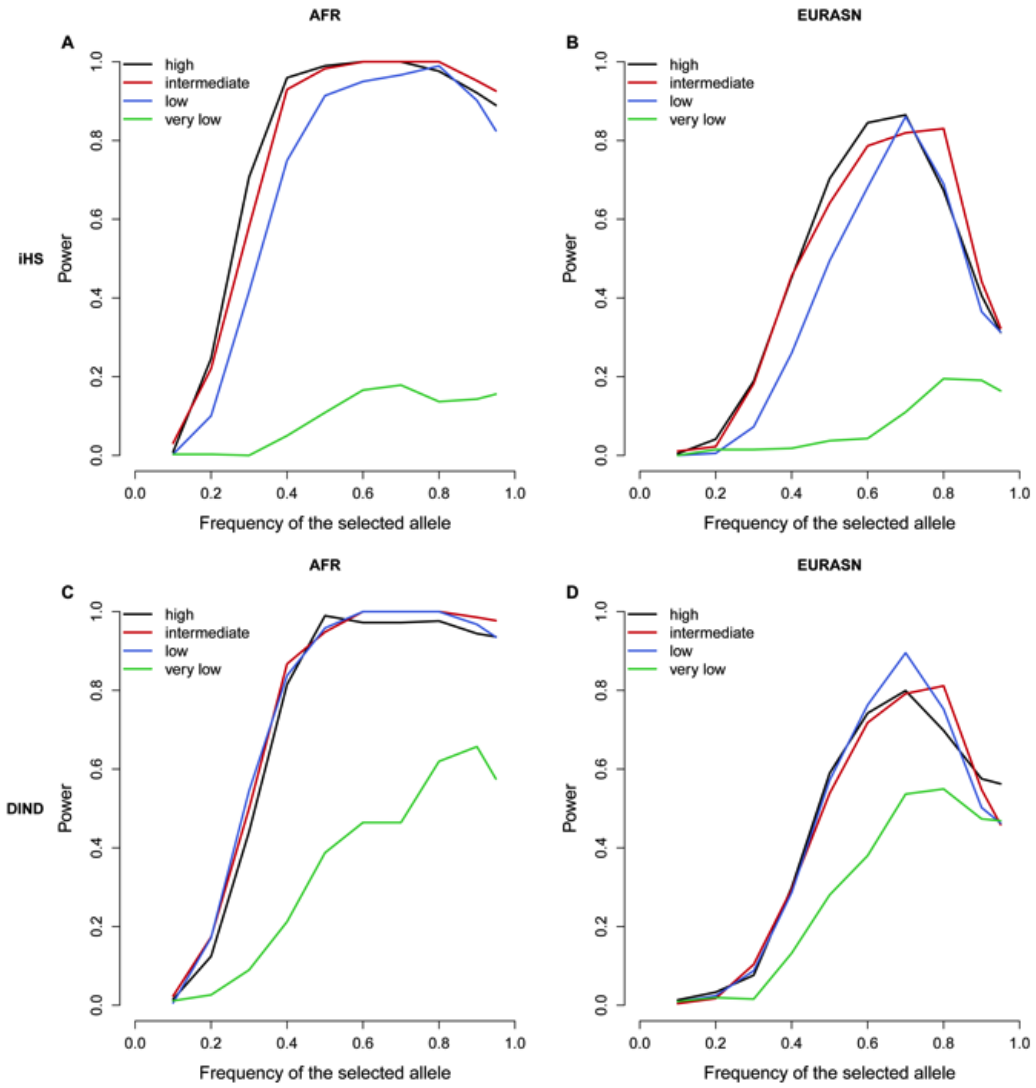


FIG. S3. Power of iHS and DIND to detect hard sweeps as a function of SNP density. We simulated "full sequence data" (100 kb) under recent positive selection ($2Ns$=100) using various mutation rate (θ=$4N\mu$). The "intermediate" mutation rate corresponds to the results shown in figure 2 (θ=0.001, value used in all the manuscript). We obtained ~700 and ~450 mutations in average in Africans and Eurasians, respectively. With the "high" and "low" mutation rates (θ=0.002 and θ=0.0005), we obtained ~1400 and ~900 mutations, and ~350 and ~225 mutations, in Africans and Eurasians. With the "very low" mutation rate (θ=0.00005) we obtained ~35 and ~22 mutations. For the "high", "low" and "very low" cases, the critical values (FPR=0.01) were obtained from 10^3 neutral simulations performed using the corresponding mutation rate. Power of iHS (*A*-*B*) and DIND (*C*-*D*) computed using the proportion of extreme values (see Materials and Methods). In each case, we performed a total of ~2,000 simulations. (*A*, *C*) African population. (*B*, *D*) Eurasian populations (EURASN): European and Asian populations for which we used the same demographic model.

SFS_CODE command line: ./sfs_code 3 100 -N 100 -L 1 100000 -t $mut_rate -r 0.001 -TS 0.08 0 1 -Td 0.08 P 1 0.5 -Tm 0.08 P 0 1 0.2 -Tm 0.08 P 1 0 0.1 -TS 0.16 1 2 -Tm 0.16 L 0.2 0.2 0.1 0.1 0.1 0.1 -Td 0.16 P 0 50 -Tm 0.16 L 10 10 0.1 0.1 0.1 0.1 -Td 0.188 P 1 100 -Td 0.188 P 2 100 -Tm 0.188 L 10 10 10 10 10 10 100 --mutation $t_mut P $pop S 50000 G 100 -n 59 60 60 -TE 0.2

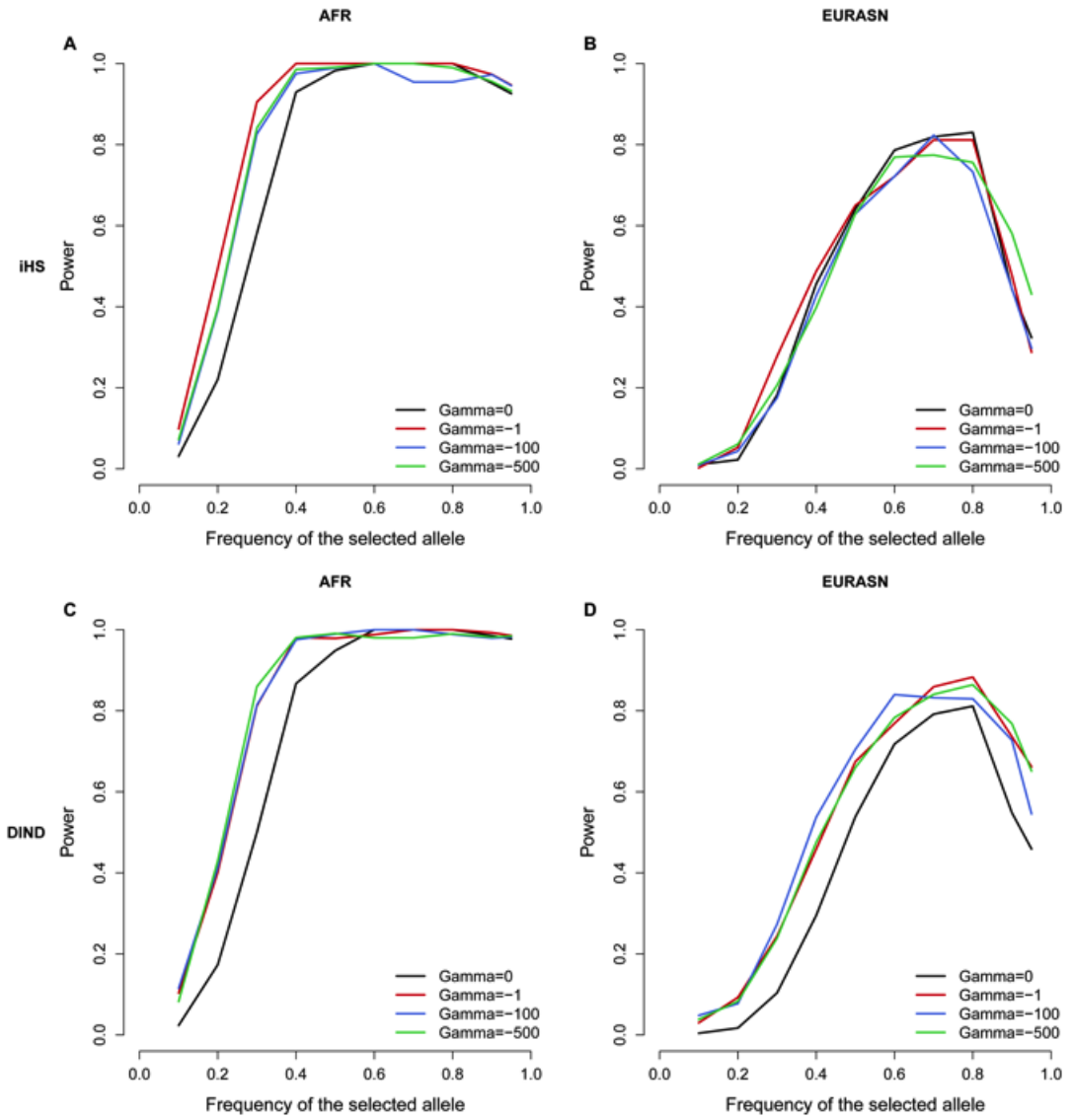


FIG. S4. Effects of background selection on the power to detect hard sweeps. We simulated the interaction between positive and background selection. We simulated "full sequence data" (100 kb) with a central SNP under recent positive selection ($2Ns$=100) while 20% of adjacent sites (evenly distributed) are under negative selection, with all sites presenting identical Gamma values (the selection parameter $2Ns$ is here termed Gamma for negatively selected sites). The Gamma values used are Gamma=-1, Gamma=-100 and Gamma=-500. The case of Gamma=0 corresponds to simulations performed under positive selection alone (fig. 2). Critical values (FPR=0.01) were obtained from 10^4 neutral simulations ($2Ns$=0, Gamma=0). Power of iHS (*A*-*B*) and DIND (*C*-*D*) computed using the proportion of extreme values (see Materials and Methods). In each case, we performed a total of ~2,000 simulations. (*A*, *C*) African population. (*B*, *D*) Eurasian populations (EURASN): European and Asian populations for which we used the same demographic model.

SFS_CODE command line: ./sfs_code 3 100 -N 100 -L 1 100000 -t 0.001 -r 0.001 -TS 0.08 0 1 -Td 0.08 P 1 0.5 -Tm 0.08 P 0 1 0.2 -Tm 0.08 P 1 0 0.1 -TS 0.16 1 2 -Tm 0.16 L 0.2 0.2 0.1 0.1 0.1 0.1 -Td 0.16 P 0 50 -Tm 0.16 L 10 10 0.1 0.1 0.1 0.1 -Td 0.188 P 1 100 -Td 0.188 P 2 100 -Tm 0.188 L 10 10 10 10 10 10 100 -n 59 60 60 -TE 0.2 --mutation $t_mut P $pop S 50000 G 100 -W 1 $Gamma 0 0.2

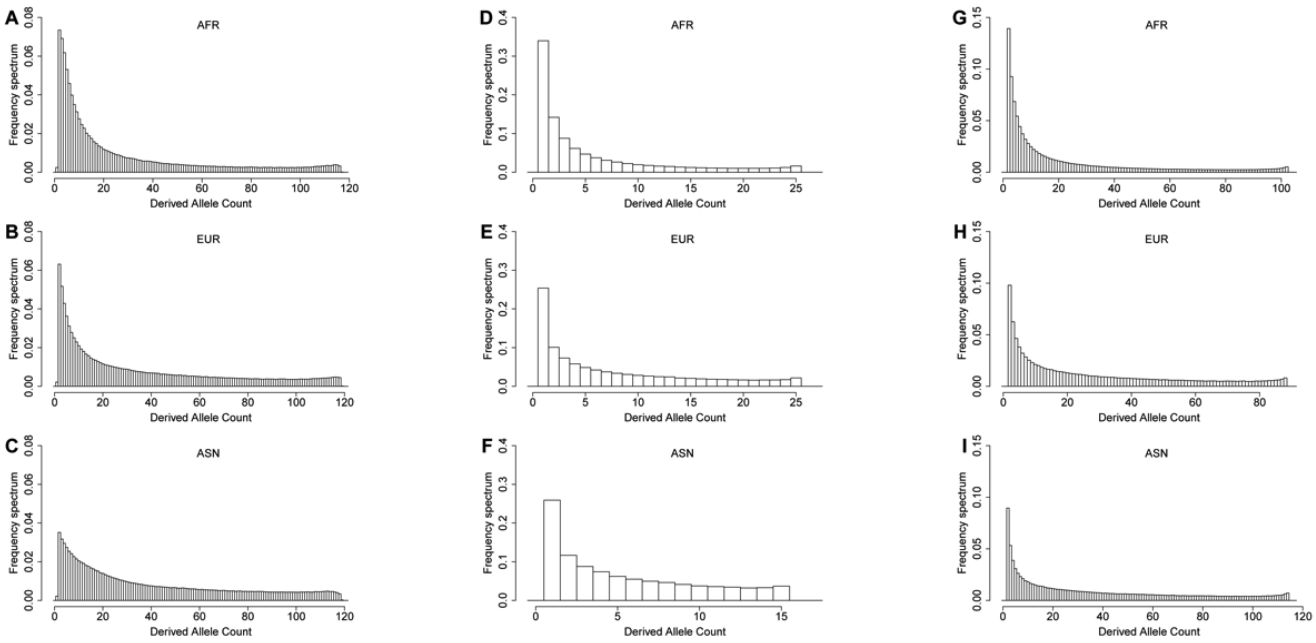


FIG. S5. Derived allele frequency spectra. (*A*-*C*) 1000 Genomes Pilot dataset. (*D*-*F*) Complete Genomics dataset. (*G*-*I*) 1000 Genomes Phase 1 dataset. We considered SNPs with unambiguously defined ancestral alleles (see Materials and Methods). AFR: African population. EUR: European population. ASN: Asian population.

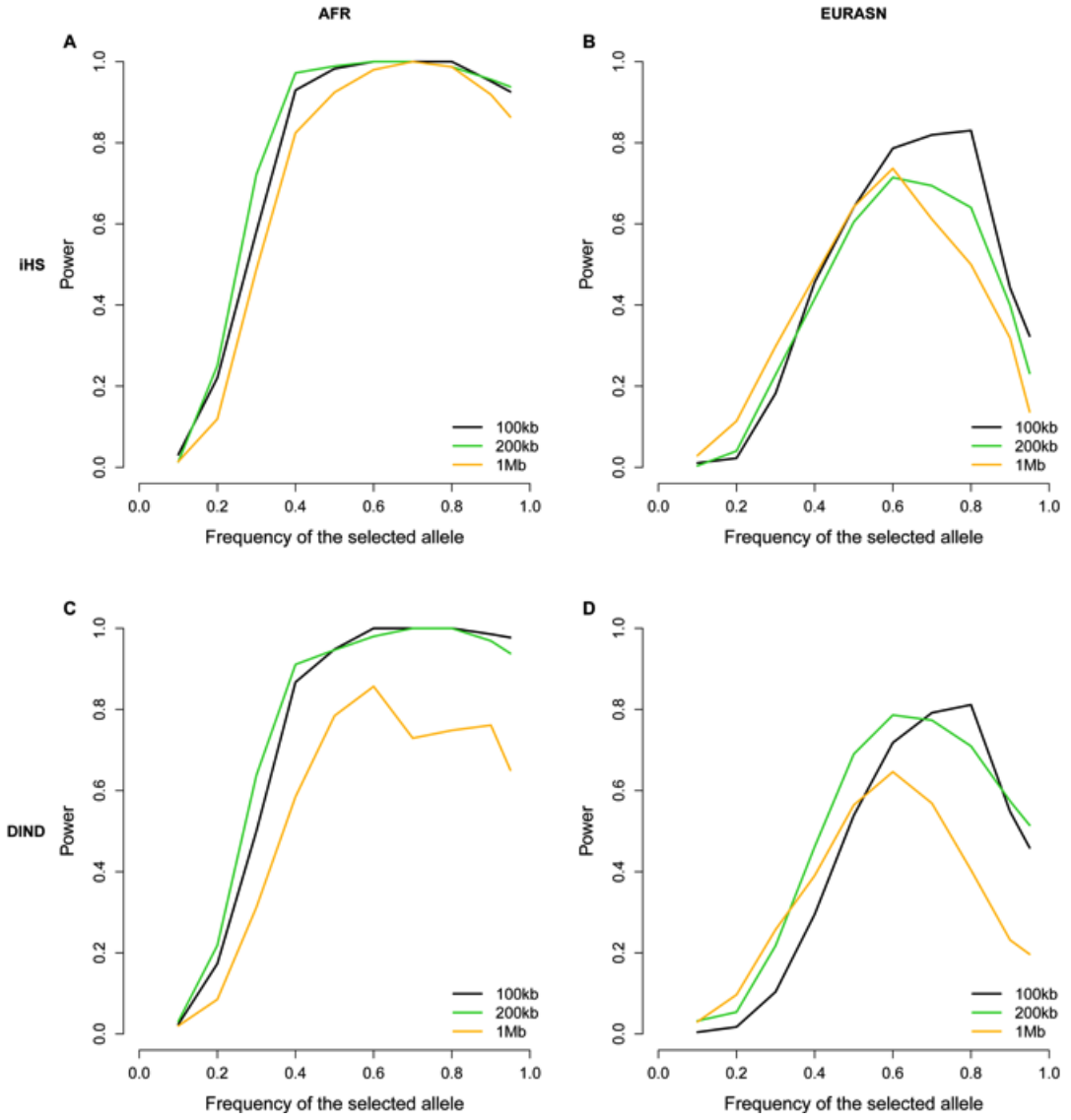


FIG. S6. Power of iHS and DIND to detect hard sweeps over regions of 100 kb, 200 kb and 1 Mb. We simulated "full sequence data" with a SNP under recent positive selection ($2Ns$=100, see Materials and Methods) inserted into the middle of each 200 kb and 1 Mb region. The case 100 kb corresponds to the results shown in figure 2. For the 200 kb and 1 Mb cases, the critical values (FPR=0.01) were obtained from 10^3 neutral simulations using the corresponding length. Power of iHS (*A*-*B*) and DIND (*C*-*D*) computed using the proportion of extreme values (see Materials and Methods). In each case, we performed a total of ~2,000 simulations. (*A*, *C*) African population. (*B*, *D*) Eurasian populations (EURASN): European and Asian populations for which we used the same demographic model.

SFS_CODE command line: ./sfs_code 3 100 -N 100 -L 1 $region_length -t 0.001 -r 0.001 -TS 0.08 0 1 -Td 0.08 P 1 0.5 -Tm 0.08 P 0 1 0.2 -Tm 0.08 P 1 0 0.1 -TS 0.16 1 2 -Tm 0.16 L 0.2 0.2 0.1 0.1 0.1 0.1 -Td 0.16 P 0 50 -Tm 0.16 L 10 10 0.1 0.1 0.1 0.1 -Td 0.188 P 1 100 -Td 0.188 P 2 100 -Tm 0.188 L 10 10 10 10 10 10 100 -n 59 60 60 -TE 0.2 --mutation $t_mut P $pop S $region_length/2 G 100

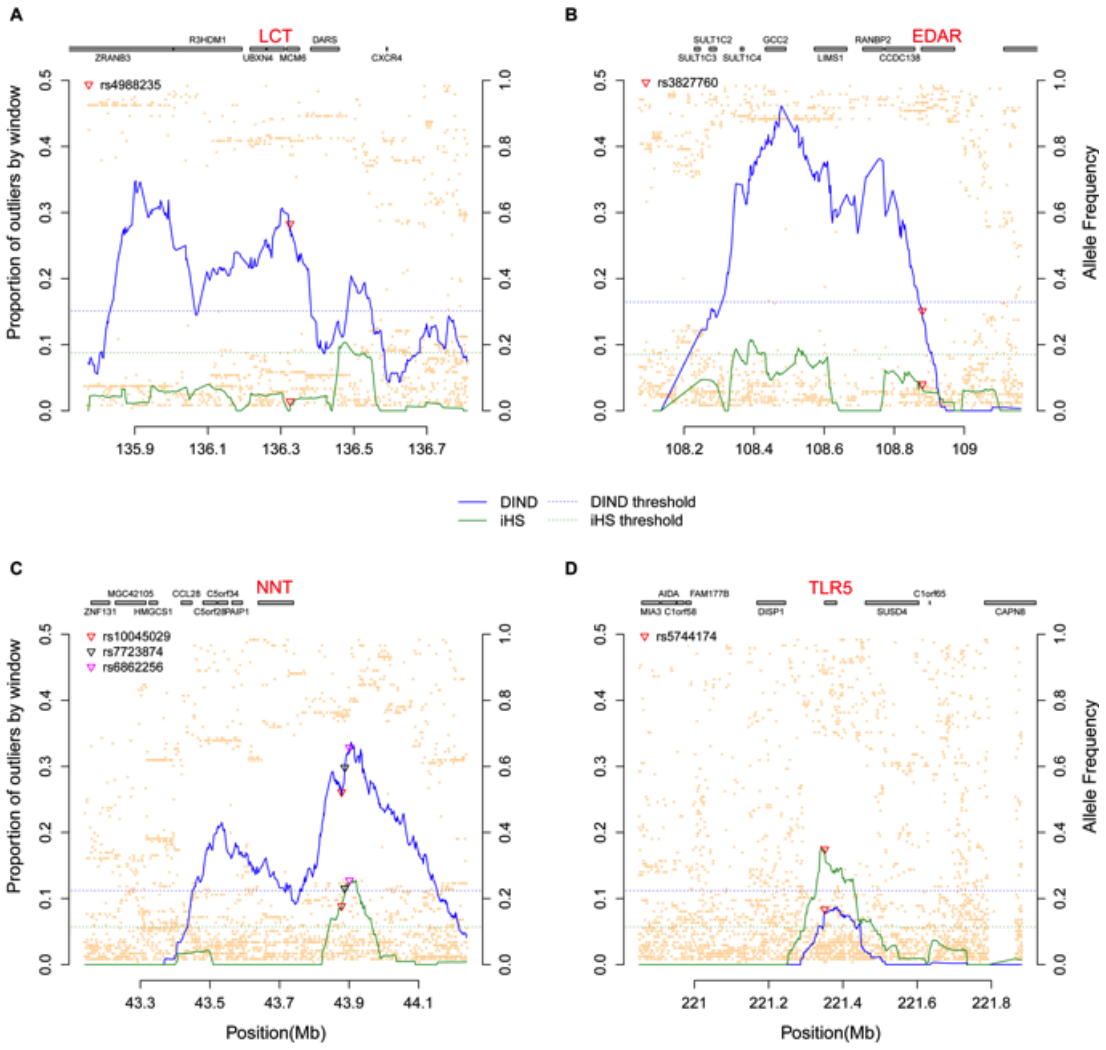


FIG. S7. Examples of genic regions under positive selection in 1000 Genomes Pilot data. DIND and iHS were computed on 1 Mb windows. Lines show the proportion of iHS (in green) and DIND (in blue) outliers by windows. The dotted lines represent, for iHS and DIND, the threshold defining the 1% most extreme proportions of outliers by window (100 kb). The orange dots are the derived allele frequencies. The gray rectangles show the position of the genes. (*A*) *LCT*. Evidence of positive selection in the EUR population at locus 2q21, centered on SNP rs4988235, responsible for lactase persistence in adulthood (red triangle). (*B*) *EDAR*. Evidence of positive selection in the ASN population at locus 2q13, around the SNP rs3827760, associated with hair morphology (red triangle). (*C*) *NNT*. Evidence of positive selection in the AFR population at locus 5p12, implicated in familial glucocorticoid and cortisol deficiency, and particularly around the SNPs rs10045029, rs7723874 and rs6862256, associated with *NNT* expression (red, dark blue and magenta triangles, respectively). (*D*) *TLR5* region. Evidence of positive selection in the AFR population at locus 10q24, involved in the recognition of bacterial flagellin, and, in particular, around SNP rs5744174, a non-synonymous mutation (L616F) associated with lower levels of NF-kB signaling in response to flagellin (red triangle).

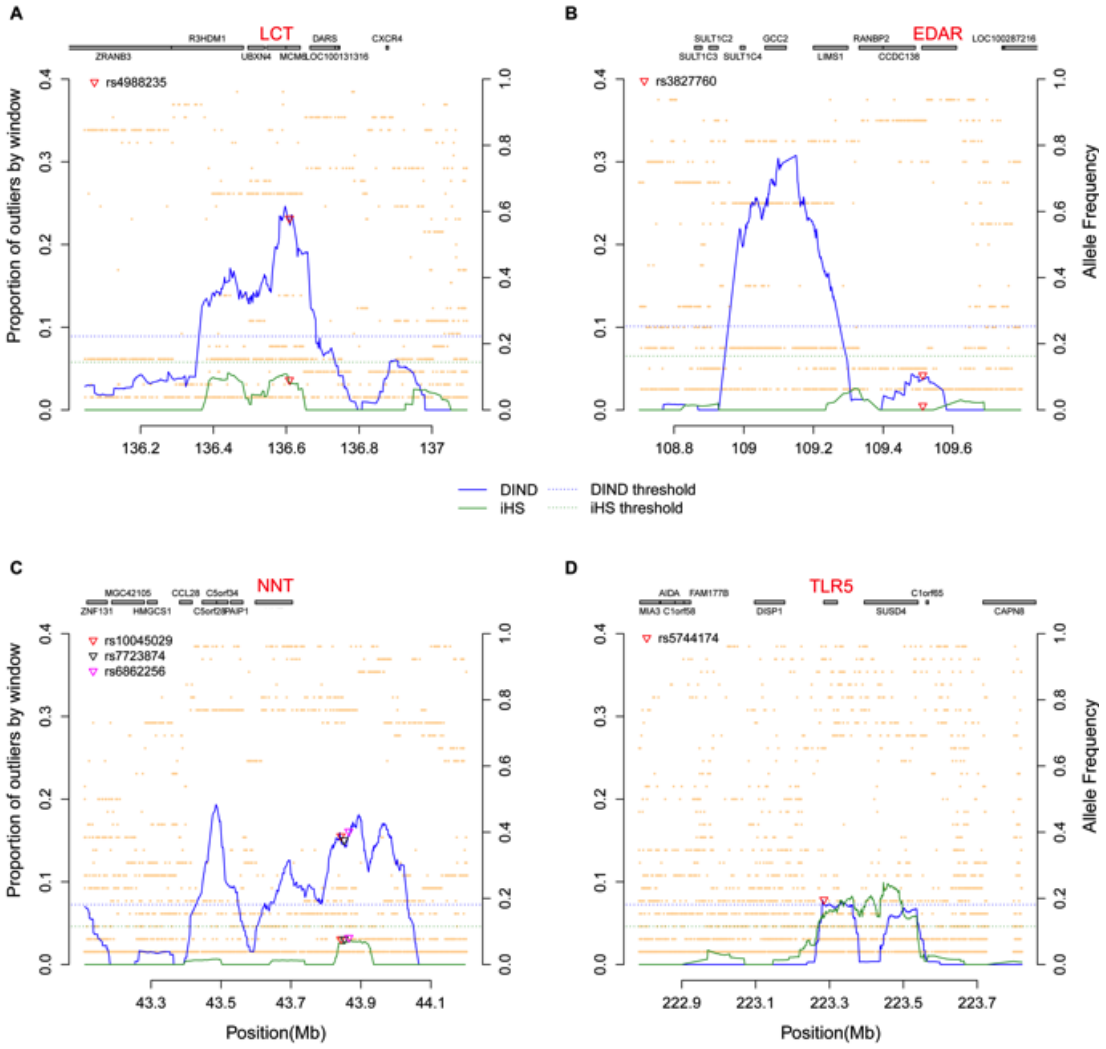


FIG. S8. Examples of genic regions under positive selection in the Complete Genomics dataset. DIND and iHS were computed on 100 kb windows. Lines show the proportion of iHS (in green) and DIND (in blue) outliers by windows. See supplementary fig. S7, Supplementary Material online, for legend explanations.

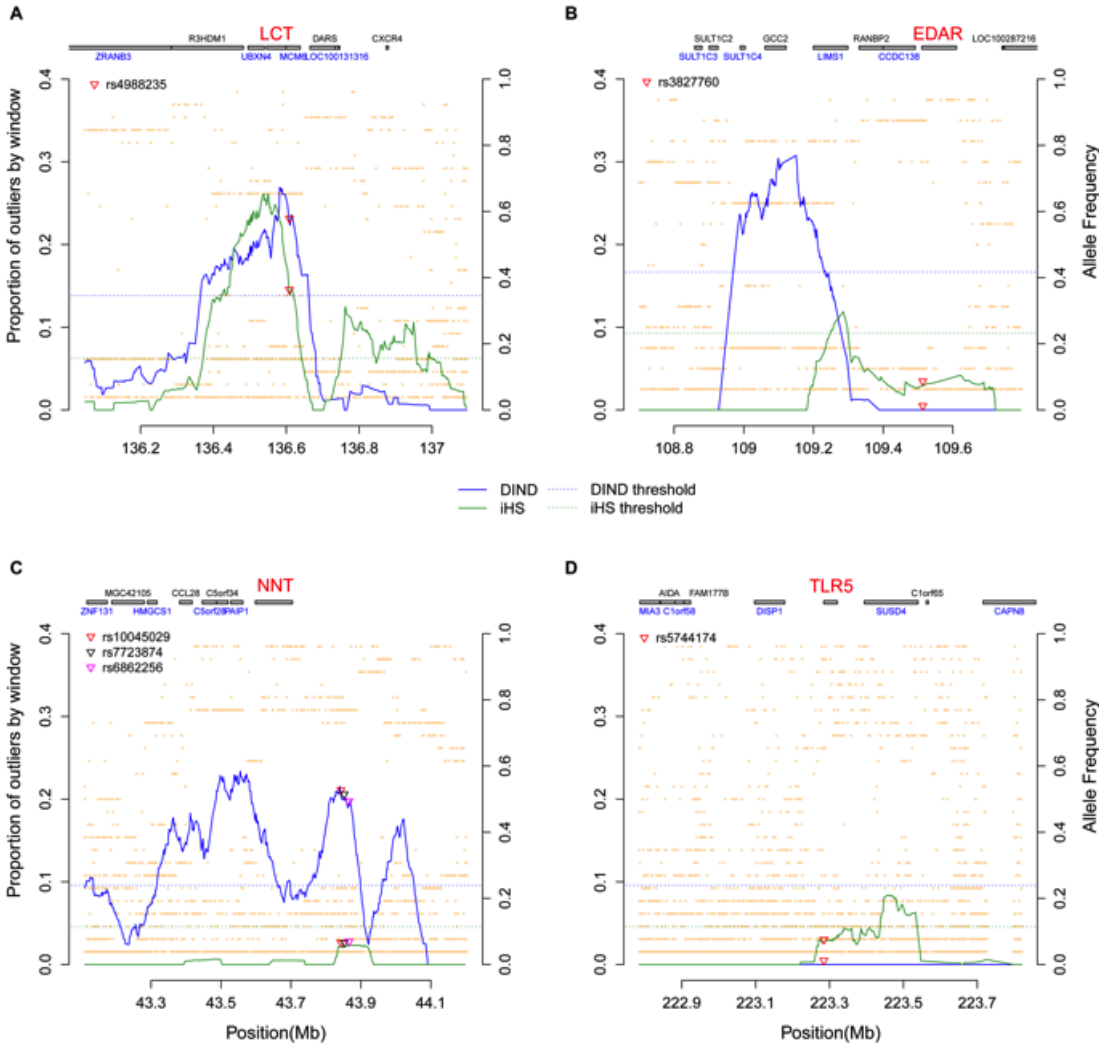


FIG. S9. Examples of genic regions under positive selection in the Complete Genomics dataset. DIND and iHS were computed on 1 Mb windows. Lines show the proportion of iHS (in green) and DIND (in blue) outliers by windows. See supplementary fig. S7, Supplementary Material online, for legend explanations.

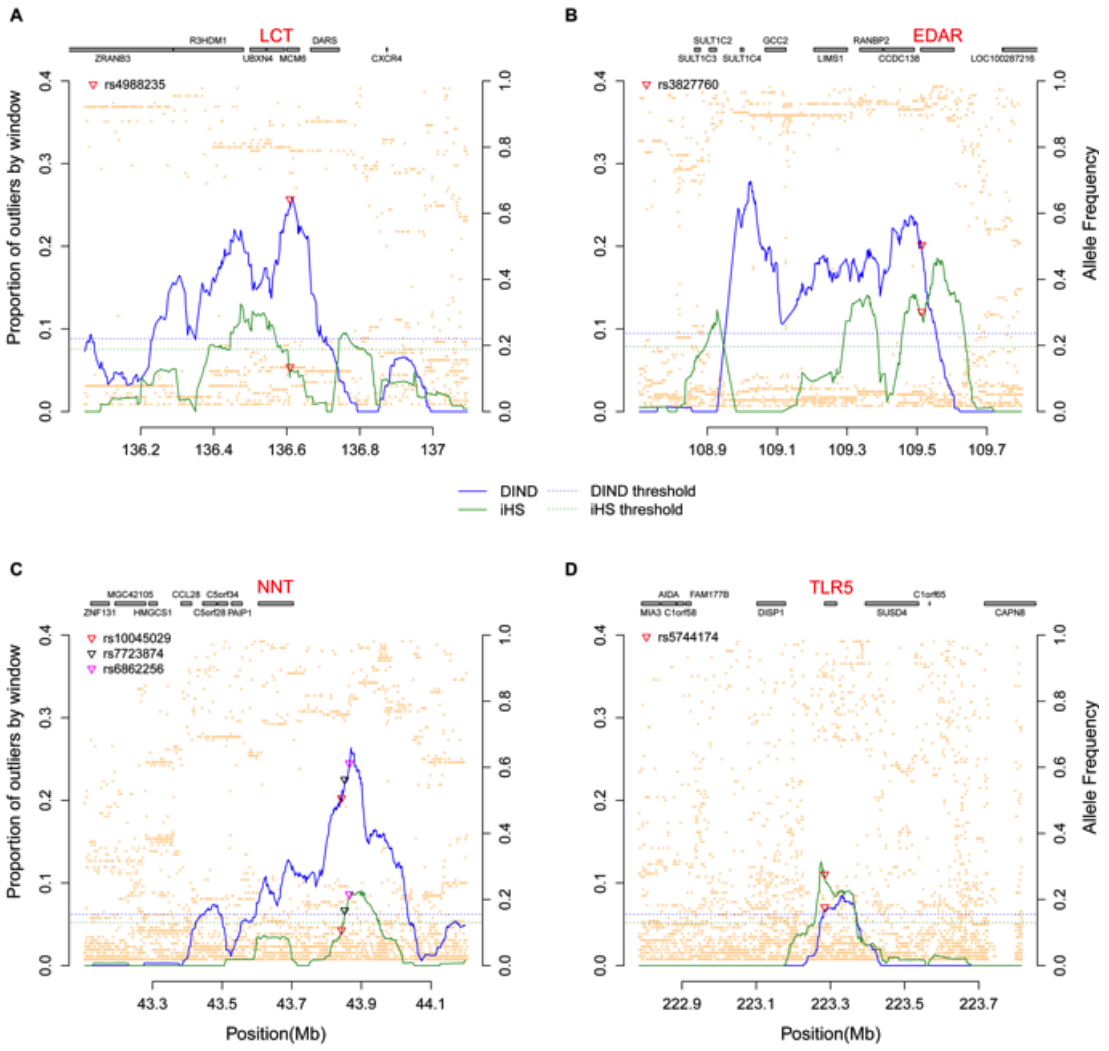


FIG. S10. Examples of genic regions under positive selection in the 1000 Genomes Phase 1 dataset. DIND and iHS were computed on 100 kb windows. Lines show the proportion of iHS (in green) and DIND (in blue) outliers by windows. See supplementary fig. S7, Supplementary Material online, for legend explanations.

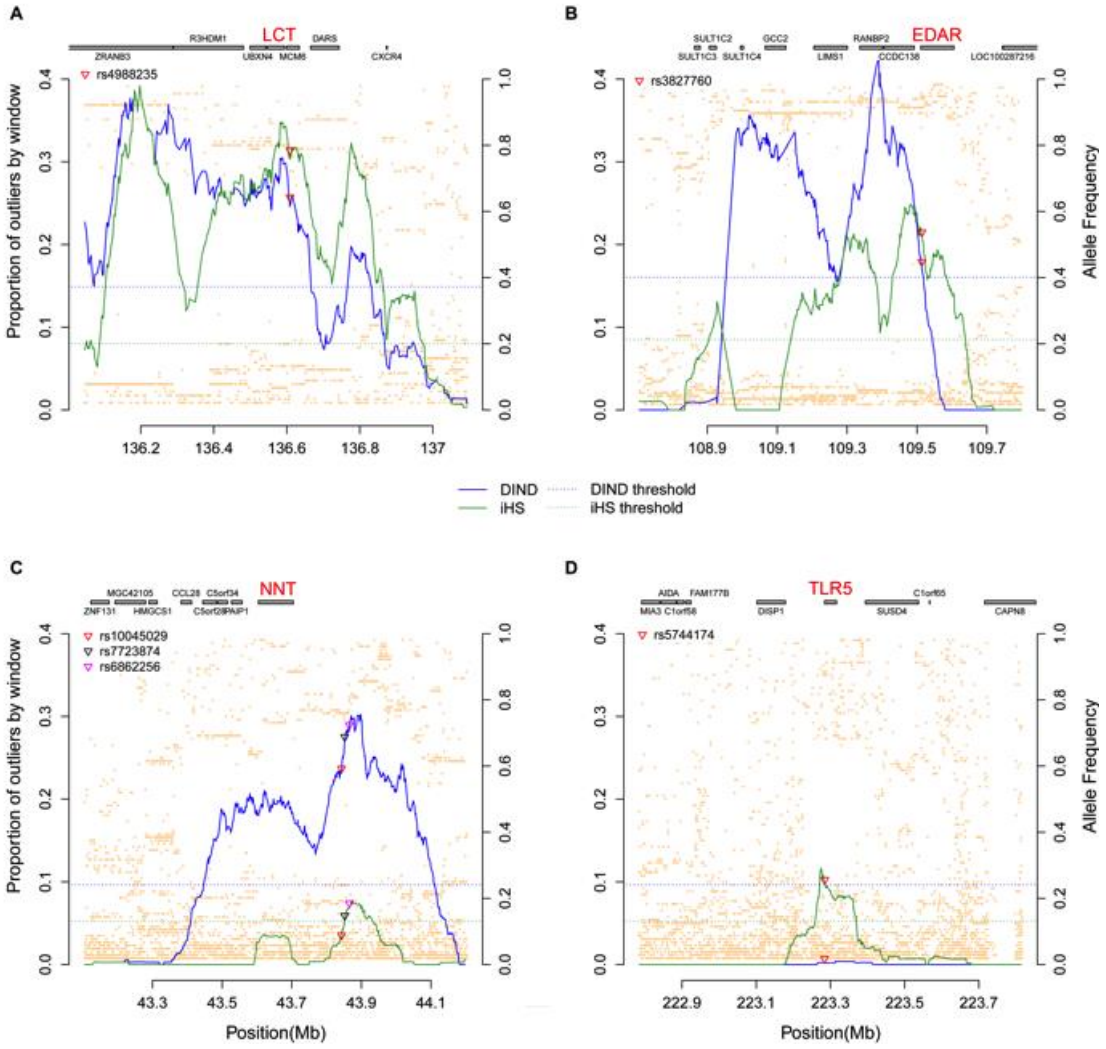


FIG. S11. Example of genic regions under positive selection in the 1000 Genomes Phase 1 dataset. DIND and iHS were computed on 1 Mb windows. Lines show the proportion of iHS (in green) and DIND (in blue) outliers by windows. See supplementary fig. S7, Supplementary Material online, for legend explanations.

Table S1. Simulated models and their corresponding posterior probabilities.

B[a]	Exp1[b]	Exp2[c]	Freq[d]	Num[e]	B[a]	Exp1[b]	Exp2[c]	Freq[d]	Num[e]
	1	1	0.000	0		1	1	0.000	0
	1	50	0.000	0		1	50	0.000	0
	1	100	0.000	0		1	100	0.000	0
	50	1	0.000	0		50	1	0.039	1579
0.1	50	50	0.000	0	**0.6**	50	50	0.040	1598
	50	100	0.000	0		**50**	**100**	**0.049**	**1946**
	100	1	0.000	0		100	1	0.032	1286
	100	50	0.000	0		100	50	0.038	1512
	100	100	0.000	0		100	100	0.033	1331
	1	1	0.000	0		1	1	0.000	0
	1	50	0.000	0		1	50	0.000	0
	1	100	0.000	0		1	100	0.000	0
	50	1	0.000	0		50	1	0.040	1617
0.2	50	50	0.000	3	0.7	50	50	0.041	1653
	50	100	0.000	2		50	100	0.027	1095
	100	1	0.000	0		100	1	0.044	1763
	100	50	0.000	0		100	50	0.030	1182
	100	100	0.000	2		100	100	0.031	1236
	1	1	0.000	0		1	1	0.000	0
	1	50	0.000	0		1	50	0.000	0
	1	100	0.000	0		1	100	0.000	0
	50	1	0.000	0		50	1	0.035	1405
0.3	50	50	0.003	138	0.8	50	50	0.014	542
	50	100	0.005	194		50	100	0.013	534
	100	1	0.000	1		100	1	0.037	1496
	100	50	0.005	181		100	50	0.011	421
	100	100	0.005	189		100	100	0.014	554
	1	1	0.000	0		1	1	0.000	0
	1	50	0.000	0		1	50	0.000	0
	1	100	0.000	0		1	100	0.000	0
	50	1	0.001	54		50	1	0.017	686
0.4	50	50	0.027	1068	0.9	50	50	0.007	264
	50	100	0.031	1256		50	100	0.009	376
	100	1	0.001	52		100	1	0.019	772
	100	50	0.030	1206		100	50	0.007	289
	100	100	0.024	946		100	100	0.010	389
	1	1	0.000	0		1	1	0.000	0
	1	50	0.000	0		1	50	0.000	0
	1	100	0.000	0		1	100	0.000	0
	50	1	0.016	656		50	1	0.005	214
0.5	50	50	0.043	1726	1	50	50	0.004	140
	50	**100**	**0.049**	**1946**		50	100	0.003	106
	100	1	0.011	441		100	1	0.011	449
	100	50	0.039	1563		100	50	0.003	121
	100	100	0.043	1711		100	100	0.003	109

[a] Intensity of the bottleneck outside Africa (e.g., 0.5 corresponds to a reduction by a half of the population size before the bottleneck).

[b] Instantaneous expansion in Africa (e.g., 50 corresponds to an increase of population size by a factor of 50) starting 20,000 years ago. Models involving more ancient

instantaneous expansions were also simulated (75,000 and 150,000 years ago) but their posterior probabilities (all equal to 0) are not reported.

[c] Instantaneous expansions in Europe and Asia (e.g., 100 corresponds to an increase of population size by a factor of 100).

[d] Estimates of the posterior probability of each model, i.e., the number of retained simulations for the model considered divided by the total number of retained simulations (see supplementary text, Supplementary Material online). The models with the highest probability are indicated in bold.

[e] Number of retained simulations for the model considered.

SFS_CODE command line: ./sfs_code 3 100 -N 100 -L 1 1300 -t 0.001 -r 0.001 -TS 0.08 0 1 -Td 0.08 P 1 $B -Tm 0.08 P 0 1 0.2 -Tm 0.08 P 1 0 0.1 -TS 0.16 1 2 -Tm 0.16 L 0.2 0.2 0.1 0.1 0.1 0.1 -Td $t_Exp1 P 0 $Exp1 -Tm 0.16 L 10 10 0.1 0.1 0.1 0.1 -Td 0.188 P 1 $Exp2 -Td 0.188 P 2 $Exp2 -Tm 0.188 L 10 10 10 10 10 10 100 -n 95 60 60 -TE 0.2

The values used for $B, $Exp1 and $Exp2 are given in the table. "$t_Exp1" is the onset of the instantaneous expansion in Africa ($t_Exp1 is equal to 0.16 for the values shown in the table)

Table S2. Power to detect hard sweeps in the context of full sequence data (2,500 simulations setting Ns=100, see Methods).

	$0 \leq SAF^{e} < 0.2$		$SAF^{e} \geq 0.2$	
Test	**AFR**	**EURASN[f]**	**AFR**	**EURASN[f]**
Tajima's D[a,b]	0.84%	0.64%	27.66%	25.86%
Fu and Li's D[a,b]	0.59%	0.85%	29.79%	12.07%
Fu and Li's F[a,b]	0.42%	0.64%	33.87%	19.63%
Fay and Wu's H[a,b]	1.17%	1.17%	42.20%	31.43%
DH[b, c]	1.09%	0.96%	37.23%	31.70%
DIND[d]	2.35%	0.43%	76.06%	38.20%
iHS[d]	3.19%	1.06%	78.90%	40.98%

[a] AFS-based statistics.
[b] The power is the proportion of simulated regions with a value of AFS-based statistics among the 1% most extreme values obtained using 10^4 neutral simulations.
[c] DH is a combined statistics based on both Tajima's D and Fay and Wu's H.
[d] For the computation of power, see Materials and Methods and the legends of figs. 2 and 3.
[e] Selected Allele Frequencies.
[f] Results for the European and Asian populations, for which we used the same demographic model.

Table S3. Power of iHS and DIND to detect background selection in the context of full sequenced data (2,500 simulations for each *2Ns* values).

	iHS		DIND	
2Ns	**AFR[a]**	**EURASN[b]**	**AFR[a]**	**EURASN[b]**
-1	2.0%	1.5%	0.0%	0.5%
-5	1.0%	0.5%	1.5%	0.5%
-10	1.0%	1.0%	2.0%	2.5%
-50	2.0%	1.0%	2.0%	0.5%
-100	0.0%	1.5%	0.5%	2.0%
-500	1.0%	0.0%	1.5%	1.5%
Average	1.17%	0.92%	1.25%	1.25%

[a]Results for the African population.
[b]Results for the European and Asian populations, for which we used the same demographic model.

Table S4. Power of iHS and DIND to detect positive selection when the coefficient of selection is low (2,500 simulations setting 2*Ns*=5).

Simulations	Test[a]	AFR	EUR[b]	ASN[b]
Full sequence data	DIND	1.32%	0.86%	0.86%
	iHS	0.82%	1.00%	1.00%
Low coverage phased[c] (1000 Genomes)	DIND	1.40%	0.88%	0.90%
	iHS	0.93%	0.82%	0.90%
Small sample size phased[d] (Complete Genomics)	DIND	1.72%	0.82%	1.27%
	iHS	1.29%	1.51%	0.99%

[a]The thresholds were computed using the proportion of outliers by 100 kb sequences (see Materials and Methods).
[b]The power of the statistics are identical for EUR and ASN because we used the same demographic model for both populations.
[c]The coverage is equal to 4× for AFR, 5× for EUR, and 3× for ASN.
[d]The sample size is equal to 13 for AFR and EUR, and 8 for ASN.

Table S5. Power of iHS and DIND to detect selection on standing variation (1,000 simulations for each standing variation scenario, see Materials and Methods).

			Current frequency of the selected allele				
2*Ns*	**Test**	**Standing variation[a]**	**0.2-0.4**	**0.4-0.6**	**0.6-0.8**	**0.8-0.1**	**Average[b]**
100	DIND	≤0.1	8.33%	77.78%	91.67%	70.83%	62.15%
		≥0.2	0.00%	9.72%	26.60%	24.09%	15.10%
		ratio	-	8	3.45	2.94	4.12
	iHS	≤0.1	0.00%	22.22%	83.33%	36.11%	35.42%
		≥0.2	0.00%	9.72%	15.96%	8.76%	8.61%
		ratio	-	2.29	5.22	4.12	4.11
1000	DIND	≤0.1	7.69%	62.50%	73.33%	63.64%	51.79%
		≥0.2	0.00%	1.43%	5.88%	22.12%	7.36%
		ratio	-	43.75	12.47	2.88	7.04
	iHS	≤0.1	30.77%	62.50%	86.67%	45.45%	56.35%
		≥0.2	2.63%	4.29%	11.76%	8.65%	6.83%
		ratio	11.69	14.58	7.37	5.25	8.25

[a]Initial frequency of the selected allele. The ratio is the power to detect selection when the initial frequency of the selected allele is ≤0.1 (case including the hard sweep model) divided the power to detect selection in the corresponding bin of initial frequency of the selected allele is ≥0.2.
[b]Values averaged between the different bins of the current frequency of the selected allele

Table S6. Description of the Complete Genomics and 1000G Phase 1 data used in this study.

Dataset	Pool	Population[a]	Coriell ID
Complete Genomics	AFR	YRI	NA18501[b], NA18502[b], NA18504[b], NA18505[b], NA18508[b], NA18517[b], NA19129[b], NA19238[cd], NA19239[cd], NA19240[c]
		LWK	NA19017, NA19020, NA19025, NA19026
	EUR	CEU	NA06985[b], NA06994[b], NA07357[b], NA10851[b], NA12004[b], NA12889c[d], NA12890[cd], NA12891[cd], NA12892[cd], NA12877[c], NA12878[c], NA12879[c], NA12880[c], NA12881[c], NA12882[c], NA12883[c], NA12884[c], NA12885[c], NA12886[c], NA12887[c], NA12888[c], NA12893[c]
		TSI	NA20502, NA20509, NA20510, NA20511
	ASN	CHB	NA18526[b], NA18537[b], NA18555[b], NA18558[b]
		JPT	NA18940[b], NA18942[b], NA18947[b], NA18956[b]
1000G Phase 1[e]	AFR	YRI	NA18486, NA18489, NA18498, NA18499, NA18501, NA18502, NA18504, NA18505, NA18507, NA18508, NA18510, NA18511, NA18516, NA18517, NA18519, NA18520, NA18522, NA18523, NA18853, NA18856, NA18858, NA18861, NA18870, NA18871, NA18907, NA18909, NA18912, NA18916, NA19093, NA19098, NA19099, NA19102, NA19108, NA19114, NA19116, NA19119, NA19129, NA19131, NA19137, NA19138, NA19147, NA19152, NA19160, NA19171, NA19172, NA19190, NA19200, NA19204, NA19207, NA19209, NA19225, NA19257
	EUR	CEU	NA06986, NA06994, NA07000, NA07037, NA07051, NA07347, NA07357, NA10847, NA10851, NA11829, NA11830, NA11831, NA11894, NA11919, NA11920, NA11931, NA11992, NA11993, NA11994, NA11995, NA12003, NA12004, NA12006, NA12043, NA12044, NA12045, NA12144, NA12154, NA12155, NA12249, NA12287, NA12489, NA12716, NA12717, NA12749, NA12750, NA12751, NA12761, NA12763, NA12812, NA12814, NA12815, NA12872, NA12873, NA12874
	ASN	CHB	NA18526, NA18532, NA18537, NA18542, NA18545, NA18547, NA18550, NA18552, NA18555, NA18558, NA18561, NA18562, NA18563, NA18564, NA18566, NA18570, NA18571, NA18572, NA18573, NA18576, NA18577, NA18579, NA18582, NA18592, NA18593, NA18603, NA18605, NA18608, NA18609, NA18631, NA18634, NA18638
		JPT	NA18940, NA18942, NA18943, NA18944, NA18945, NA18947, NA18948, NA18949, NA18951, NA18952, NA18953, NA18956, NA18959, NA18960, NA18961, NA18964, NA18965, NA18968, NA18971, NA18973, NA18974, NA18975, NA18976, NA18980, NA18981, NA19005

[a]The samples come from HapMap populations: YRI: Yoruba in Ibadan, Nigeria; LWK: Luhya in Webuye, Kenya; CEU: Utah residents with Northern and Western European ancestry; TSI: Toscans in Italy; CHB: Han Chinese in Beijing, China; JPT: Japanese in Tokyo, Japan.
[b]Samples that are present both in the 1000G Pilot dataset and the Complete Genomics dataset.
[c]These samples are members of a father-mother-daughter trio (Africans), or of a pedigree with the four grand-parents, the two parents and the 11 children (European). All the samples are used for the phasing process. Only founders are kept for positive selection analysis.
[d]Unrelated founders from the African trio and the European pedigree.
[e]All these samples are present in the 1000 Genome Pilot dataset. For a complete list of the samples from 1000G Pilot dataset, see (The 1000 Genomes Project Consortium 2010).

Table S7. Number of genic and non-synonymous SNPs with a clearly defined ancestral allele together with a derived allele frequency (DAF) >0.2 in each population.

	1000G Pilot[a]			Complete Genomics[b]			1000G Phase 1[a]		
Functional Classes	**AFR**	**EUR**	**ASN**	**AFR**	**EUR**	**ASN**	**AFR**	**EUR**	**ASN**
Total	3,033,113	2,912,956	2,727,770	2,945,000	2,601,829	2,465,279	3,685,407	3,358,220	3,199,202
Genic	1,119,635	1,071,872	1,011,409	1,118,362	1,001,526	942,725	1,346,645	1,228,561	1,169,973
Non-synonymous	7,450	7,291	7,461	6,916	7,859	6,274	7,976	7,724	7,297

[a] The dataset was annotated using GENCODE gene models and the Human Genome Mutation Database (see The 1000 Genomes Project Consortium 2010, Supplementary Information 13.2)

[b] The dataset was annotated using NCBI annotation build 37.2 (see Complete Genomics public genomes information, http://www.completegenomics.com/public-data/69-Genomes/)

Table S8. Enrichment in genic vs. non-genic among iHS outliers computed on HapMap datasets.

Population	Voight et *al.*[a]	iHS[b]
AFR	**1.23*****	**1.25****
EUR	**1.16*****	**1.17***
ASN	**1.13*****	1.08

*P<0.01; **P<0.001; ***P<10^{-20}

[a] Analysis of HapMap phase 1 dataset (Voight et al. 2006).

[b] Analysis of HapMap phase 2 dataset with P-values obtained following the resampling method used in this study (see Materials and Methods).

Table S9. Enrichment in genic vs non-genic and non-synonymous vs non-genic SNPs among iHS and DIND outliers computed from windows of 1 Mb.

		Genic vs. Non-genic						**Non-synonymous vs. Non-genic**					
		1000G Pilot		**Complete Genomics**		**1000G Phase 1**		**1000G Pilot**		**Complete Genomics**		**1000G Phase 1**	
Population	**Odds ratio (OR)**	**DIND**	**iHS**	**DIND**	**iHS**	**DIND**	**iHS**	**DIND**	**iHS**	**DIND**	**iHS**	**DIND**	**iHS**
AFR	OR	**1.48****	0.77	**1.41****	0.98	**1.46****	0.88	**1.56****	0.67	**1.66****	0.83	**1.32***	0.73
	OR_C	**1.51****	0.81	**1.24****	0.99	**1.45****	0.89	**1.53****	0.71	**1.39***	0.82	**1.24°**	0.70
EUR	OR	**1.32****	0.84	**1.26****	1.02	**1.38****	1.01	**1.37***	0.98	**1.53****	1.13	**1.48****	1.09
	OR_C	**1.47****	0.95	**1.24****	1.06	**1.42****	1.06	**1.58****	1.07	**1.52****	1.18	**1.61****	1.09
ASN	OR	1.13	0.80	1.00	1.03	**1.20***	0.89	1.05	0.69	1.23	1.07	1.18	0.87
	OR_C	**1.30****	0.93	1.02	1.08	**1.47****	0.94	**1.41***	0.85	**1.29***	1.10	**1.72****	0.88

°P<0.05; *P<0.01; **P<0.001

OR_C indicates that the logistic regression used to compute the odds ratio controlled for the following confounding factors: mean recombination rate, mean coverage and number of SNPs by windows.

Table S13. Enrichment in GWAS-SNPs among DIND outliers measured on the basis of odds ratios.

	1000G Pilot		Complete Genomics		1000G Phase1	
Population	**100 kb[a]**	**1 Mb[b]**	**100 kb[a]**	**1 Mb[b]**	**100 kb[a]**	**1 Mb[b]**
AFR	1.15	1.15	0.97	1.15	1.25	0.93
EUR	1.56	2.25	1.49	2.16	1.77	1.92
ASN	0.80	0.71	0.49	0.59	0.97	0.62

[a] DIND values were computed using 100 kb windows (fig. 5).
[b] DIND values were computed using 1 Mb windows, but the proportion of SNPs to extract list of candidate genes were computed using 100 kb windows.

Table S14. Enrichment in SNPs in genome-wide association among DIND outliers, computed per trait and disease and for windows of 100 kb and 1 Mb.

		100kb windows			1Mb windows		
	Traits or diseases	**1000G Pilot**	**CG**	**1000G Phase 1**	**1000G Pilot**	**CG**	**1000G Phase 1**
AFR - DIND	Alcoholism	99.000	-	-	99.000	-	33.000
	Adiponectin levels	-	-	-	11.000	-	-
	Blood pressure	-	-	-	-	6.600	5.211
	Body mass	-	-	4.304	-	-	-
	Bone mineral density	14.143	-	-	-	-	-
	Breast cancer	-	-	-	7.615	-	8.250
	Cholesterol	-	1.833	3.536	3.882	-	1.737
	Crohn's disease	2.357	-	-	4.829	-	-
	Dupuytren's disease	-	33.000	-	-	-	-
	Gamma glutamyl transpeptidase	-	-	-	-	99.000	-
	Height	2.571	2.640	3.908	1.269	1.375	2.605
	Leukemia	-	-	-	24.750	-	-
	Menarche (age at onset)	-	-	-	4.500	10.421	8.609
	Multiple sclerosis	3.667	-	3.667	-	-	-
	Pancreatic cancer	49.500	99.000	49.500	-	-	-
	Primary biliary cirrhosis	9.900	11.000	12.375	-	-	-
	Psychiatric diseases	16.500	-	18.000	7.615	-	8.250
	QT interval	-	16.500	-	-	16.500	-
	Triglycerides	-	-	-	-	19.800	-
	Type 2 diabetes	3.536	3.808	3.194	-	-	-
	Ulcerative colitis	-	-	-	7.920	-	-
	Ventricular conduction	-	7.615	-	-	7.615	-
	All GWA-SNPs	**1.154**	**0.971**	**1.248**	**1.154**	**1.148**	**0.934**
EUR - DIND	Adiponectin levels	-	-	-	8.250	7.615	8.250
	Age-related macular degeneration	-	-	28.286	14.143	16.500	12.375
	Amyotrophic lateral sclerosis	-	-	49.500	-	-	-
	Asthma	-	8.250		-	-	-
	Blood pressure	7.333	-	3.194	7.333	7.615	6.600
	Body mass	3.000	3.094	3.094	3.000	3.000	-
	Bone mineral density	-	-	22.000	11.000	-	11.000
	Breast cancer	-	-	-	-	8.250	-
	Butyrylcholinesterase levels	-	-	-	-	49.500	-
	C-reactive protein	-	6.188	-	-	-	-
	Celiac disease	-	-	-	-	4.500	3.960
	Cholesterol	3.046	3.246	4.368	4.641	4.950	2.870
	Chronic kidney disease	6.188	6.600	5.824	-	6.600	12.375
	Coffee consumption	49.500	-	-	49.500	-	24.750
	Corneal structure	-	-	16.500	24.750	-	16.500
	Coronary heart disease	-	-	-	3.414	3.536	3.300
	Creatinine levels	-	49.500	-	-	-	-
	Crohn's disease	1.500	-	-	3.046	3.414	2.870
	Fasting glucose-related traits	-	-	-	7.071	-	7.071

	Glaucoma	-	-	-	-	24.750	24.750
	Height	2.750	4.400	0.943	0.908	2.200	2.883
	Hematological parameters	-	-	24.750	14.143	12.375	24.750
	Hemoglobin	-	-	24.750	33.000	33.000	24.750
	Immunoglobulin A	99.000	-	-	99.000	-	-
	Inflammatory bowel disease	11.000	-	-	11.000	-	-
	Keloid	-	99.000	-	-	-	-
	Menarche (age at onset)	-	-	-	3.808	-	-
	Metabolite levels	-	-	-	7.071	9.900	-
	Multiple sclerosis	3.000	3.000	-	6.188	3.000	-
	Permanent tooth development	-	-	-	49.500	33.000	49.500
	Pigmentation	5.211	-	11.000	-	-	-
	Polycystic ovary syndrome	99.000	99.000	99.000	-	-	-
	Progressive supranuclear palsy	-	-	-	24.750	39.600	-
	Proinsulin levels	-	-	-	19.800	-	16.500
	Prostate cancer	3.808	9.900	3.960	3.960	-	-
	Renal function and chronic kidney disease	-	-	99.000	-	-	-
	Retinal vascular caliber	-	-	49.500	49.500	49.500	49.500
	Rheumatoid arthritis	3.300	-	3.414	-	3.960	3.536
	Smoking behavior	-	-	-	49.500	33.000	24.750
	Testicular cancer	19.800	19.800	19.800	19.800		19.800
	Type 1 diabetes	3.000	-	6.600	6.188	6.188	6.600
	Type 2 diabetes	2.676	-	-	-	-	-
	Upper aerodigestive tract cancers	-	-	-	99.000	99.000	99.000
	Urate levels	11.000	-	9.900	-	-	-
	Vertical cup-disc ratio	-	-	-	-	24.750	16.500
	White blood cell count	-	-	24.750	24.750	33.000	24.750
	All GWA-SNPs	**1.558**	**1.490**	**1.767**	**2.247**	**2.165**	**1.921**
ASN - DIND	Alcoholism	-	-	-	24.750	-	-
	Body mass	-	-	-	-	4.125	-
	Bone mineral density	-	-	11.000	-	-	-
	Breast cancer	9.000	-	9.000	-	-	-
	Butyrylcholinesterase levels	-	-	49.500	-	-	-
	Celiac disease	5.500	5.211	-	-	-	-
	Cholesterol	-	2.020	3.414	1.833	2.020	1.678
	Crohn's disease	4.213	-	1.768	-	-	-
	Esophageal cancer	-	33.000	-	-	-	-
	Height	-	-	1.100	1.031	2.750	2.225
	Hirschsprung's disease	-	99.000	-	-	-	-
	Mean platelet volume	-	-	-	-	-	-
	Metabolite levels	-	-	-	-	9.900	-
	Myopia (pathological)	99.000	-	99.000	-	-	-
	Platelet aggregation	-	-	33.000	-	-	-
	Polycystic ovary syndrome	-	-	49.500	-	-	-
	Progressive supranuclear palsy	-	-	33.000	-	24.750	-
	Prostate cancer	3.808	4.125	-	3.960	-	-
	Psoriasis	-	-	-	-	-	7.615

	Psychiatric diseases	6.600	-	-	-	-	-
	Rheumatoid arthritis	-	-	-	3.667	-	-
	Systemic lupus erythematosus	-	-	-	8.250	-	6.600
	Testicular cancer	-	-	-	19.800	-	19.800
	Triglycerides	4.125	-	-	4.304	4.950	4.304
	Upper aerodigestive tract cancers	-	49.500	-	-	-	-
	Vertical cup-disc ratio	24.750	-	-	-	-	-
	All GWA-SNPs	**0.802**	**0.489**	**0.973**	**0.712**	**0.587**	**0.618**
AFR - iHS	Adiponectin levels	-	-	9.900	-	-	9.900
	Alcoholism	-	49.500	-	-	99.000	-
	Alzheimer's disease	-	-	19.800	-	-	9.000
	C-reactive protein	19.800	-	-	19.800	-	-
	Cholesterol	-	1.941	-	-	1.904	-
	Electrocardiographic traits	-	-	14.143	-	-	14.143
	Glycated hemoglobin levels	49.500	-	-	49.500	-	-
	Menarche (age at onset)	4.500	-	-	4.500	4.950	-
	Pancreatic cancer	-	99.000	-	-	99.000	-
	Pigmentation	-	19.800	-	-	24.750	-
	Proinsulin levels	-	-	24.750	-	-	24.750
	Rheumatoid arthritis	4.304	-	-	4.304	-	-
	Tuberculosis	-	99.000	-	-	99.000	-
	Type 2 diabetes	-	-	3.094	-	-	3.094
	Ulcerative colitis	3.808	-	3.960	3.808	-	3.960
	Ventricular conduction	-	8.250	-	-	8.250	-
	Vitamin B12 levels	-	99.000	-	-	99.000	-
	Vitiligo	16.500	-	-	16.500	-	-
	All GWA-SNPs	**0.731**	**0.698**	**0.723**	**0.731**	**0.931**	**0.619**
EUR - iHS	Asthma	-	-	-	-	8.250	-
	Blood pressure	-	-	-	-	3.808	-
	Body mass	2.912	6.188	3.000	2.912	3.094	3.000
	Bone mineral density	-	11.000	-	-	11.000	9.900
	Celiac disease	-	-	3.960	-		3.960
	Cholesterol	-	3.246	2.870	-	4.950	2.870
	Coronary heart disease	-	3.536	-	-	3.536	-
	Crohn's disease	-	-	1.414	-		1.414
	Fasting glucose-related traits	-	7.615	-	-	7.615	-
	Glycated hemoglobin levels	-	24.750	-	-	24.750	24.750
	Height	-	3.337	0.952	-	1.100	-
	Hematological parameters	14.143	-	-	14.143	-	-
	Hepatitis B	-	-	-	-	99.000	-
	Hodgkin's lymphoma	24.750	99.000	33.000	24.750	99.000	33.000
	IgE levels	24.750	49.500	49.500	24.750	49.500	49.500

	Immunoglobulin A	99.000	-	-	99.000	-	-
	Immunoglobulin A	-	-	-	-	-	-
	Inflammatory bowel disease	11.000	19.800	-	11.000	19.800	-
	Migraine	-	-	33.000	-	-	33.000
	Multiple sclerosis	3.000	3.000	-	3.000	3.000	-
	Neuroblastoma	-	-	33.000	-	-	33.000
	Osteoporosis	-	-	49.500	-	-	49.500
	Parkinson's disease	8.250	-	-	8.250	-	-
	Pigmentation	5.500	-	5.824	11.000	-	11.647
	Platelet counts	6.188	-	-	6.188	-	-
	Proinsulin levels	-	19.800	-	-	19.800	-
	Prostate cancer	-	4.714	-	-	4.714	-
	Psoriasis	6.600	7.615	6.600	6.600	7.615	6.600
	Pulmonary function	-	-	6.188	-	-	6.188
	QT interval	-	9.000	-	-	-	-
	Response to treatment (immunity)	99.000	-	-	99.000	-	-
	Systemic lupus erythematosus	-	8.250	6.188	-	8.250	6.188
	Type 2 diabetes	-	5.657	-	-	5.657	2.605
	Ulcerative colitis	-	3.000	2.750	-	3.000	-
	Vitamin B12 levels	-	-	99.000	-	-	99.000
	Vitiligo	-	-	-	-	14.143	-
	Weight	12.375	9.900	12.375	12.375	-	12.375
	All GWA-SNPs	**0.813**	**1.591**	**1.101**	**0.887**	**1.678**	**1.174**
ASN - iHS	Ankylosing spondylitis	12.375	-	-	12.375	-	-
	Blood pressure	5.500	6.188	-	5.500	13.200	-
	Body mass	3.667	-	-	3.667	-	-
	Breast cancer	-	9.000	-	-	19.800	-
	Celiac disease		-	-	-	5.824	-
	Cholesterol	-	4.714	1.650	-	4.714	1.650
	Colorectal cancer	-	11.000	-	-	-	9.900
	Crohn's disease	-	-	-	-	2.063	-
	Electrocardiographic traits	-	-	-	-	19.800	-
	Esophageal cancer	-	-	-	-	49.500	-
	Graves' disease	-	-	16.500	-	-	16.500
	Height	1.021	2.789	2.225	1.021	1.394	2.225
	Hematological parameters	-	-	-	-	14.143	-
	Hirschsprung's disease	99.000	-	99.000	99.000	-	99.000
	Hypertension	-	-	-	-	24.750	-
	Inflammatory biomarkers	-	-	-	-	49.500	-
	Leprosy	-	-	-	-	-	19.800
	Lung cancer	-	-	-	-	-	19.800
	Mean corpuscular hemoglobin	-	-	-	-	19.800	-
	Menarche (age at onset)	4.950	-	4.714	4.950	-	4.714

Menopause (age at onset)	-	-	-	-	11.000	-
Nephrolithiasis	-	-	99.000	-	-	99.000
Nephropathy	-	-	-	-	99.000	-
Permanent tooth development	-	99.000	-	-	-	-
Platelet counts	-	-	5.211	-	-	5.211
Proinsulin levels	-	-	16.500	-	-	16.500
Prostate cancer	-	-	3.960	-	-	-
Psychiatric diseases	-	9.000	-	-	-	-
Pulmonary function	-	-	-	-	6.600	-
Rheumatoid arthritis	-	-	-	-	4.714	-
Thyroid volume	33.000	-	-	-	-	-
Triglycerides	4.125	-	-	4.125	-	-
Type 1 diabetes	4.125	-	-	-	-	-
Type 2 diabetes	3.300	4.500	3.000	3.300	9.900	3.000
Ulcerative colitis	-	-	7.071	-	-	3.414
Upper aerodigestive tract cancers	-	-	-	-	49.500	-
Ventricular conduction	-	7.071	-	-	7.071	-
Vitiligo	16.500	-	-	16.500	-	-
Waist-hip ratio	-	11.000	-	-	11.000	-
Weight	-	-	-	-	33.000	-
All GWA-SNPs	**0.978**	**1.252**	**1.150**	**0.801**	**2.541**	**1.239**

[a] For comparison purpose the enrichment in SNPs in genome-wide association among iHS outliers are given.

Table S15. Number of selected alleles associated with fairer or darker skin, hair and eye pigmentation in the European population.

Population	Fairer[a]	Darker[b]	Genes	Position	References
EUR	rs1667394[d]		*OCA2*	15q	(Sulem et al. 2007)
	rs916977[d]		*HERC2*	15q13	(Kayser et al. 2008)
	rs1042602		*TYR*	11q14-q21	(Sulem et al. 2007)
Total EUR[c]	**2**	**0**			

[a]The selected allele (derived allele) is associated to fairer skin, hairs and eyes pigmentation.
[b]The selected allele (derived allele) is associated to darker skin, hairs and eyes pigmentation.
[c]The count takes into account linkage disequilibrium.
[d]These SNPs are in linkage disequilibrium and show the same direction of selection. Each set of SNPs in LD is counted one SNP.

Table S17. Number of selected alleles associated to high- or low-stature in all the populations.

Population	Increase[a]	Decrease[b]	Genes	Position	References
AFR	rs6686842		*SCMH1*	1p34	(Weedon et al. 2008)
	rs724577[d]		*LCORL*	4p15.31	(N'Diaye et al. 2011)
		rs16896068[d]	*LCORL*	4p15.31	(Weedon et al. 2008)
		rs4549631	*LOC387103*	6q22.32	(Weedon et al. 2008)
Total AFR[c]	**1**	**1**			
EUR	rs1635852		*JAZF1*	7p15.2-p15.1	(Johansson et al. 2009)
	rs9835332		*C3orf63*	3p14.3	(Lango Allen et al. 2010)
	rs724577		*LCORL*	4p15.31	(N'Diaye et al. 2011)
	rs3116602		*DLEU7*	13q14.3	(Weedon et al. 2008)
	rs9969804		*IPPK*	9q22.31	(Lango Allen et al. 2010)
		rs2093210	*SIX6*	14q23.1	(Lango Allen et al. 2010)
		rs6088813[e]	*UQCC*	20q11.22	(Soranzo et al. 2009)
		rs6060369[e]	*GDF5, UQCC, BFZB*	20q11.22	(Lettre et al. 2008)
Total EUR[c]	**5**	**2**			
ASN	rs3760318		*CRLF3, ATAD5, CENTA2, RNF135*	17q11.2	(Gudbjartsson et al. 2008)
	rs11684404		*EIF2AK3*	2p12	(Lango Allen et al. 2010)
		rs6088813	*UQCC*	20q11.22	(Soranzo et al. 2009)
	rs925098		*LCORL*	4p15.31	(Carty et al. 2012)
Total ASN[c]	3	1			

[a]The selected allele (derived allele) is associated to high-stature.
[b]The selected allele (derived allele) is associated to low-stature.
[c]The count takes into account linkage disequilibrium.
[d]rs16896068 and rs724577 are in linkage disequilibrium, and present different direction of selection. We have removed them from the total.
[e]rs6088813 and rs6060369 are in linkage disequilibrium and show the same direction of selection. Each set of SNPs in linkage disequilibrium is counted one SNP.

Table S18. Number of selected alleles associated to age of onset of menarche in all the populations.

Population	Increase[a]	Decrease[b]	Genes	Position	References
AFR	rs1361108		*C6orf173, TRMT11*	6q11.1-q22.33	(Elks et al. 2010)
	rs1364063		*NFAT5*	16q22.1	(Elks et al. 2010)
	rs6762477		*RBM6*	3p21.3	(Elks et al. 2010)
Total AFR	**3**	**0**			
EUR	rs7617480		*KLHDC8B*	3p21.31	(Elks et al. 2010)
Total EUR	**1**	**0**			

[a]The selected allele (derived allele) is associated to later menarche onset.
[b]The selected allele (derived allele) is associated to earlier menarche onset.

Table S19. Number of selected risk, protection or with non-reported effect allele for each GWAS disease categories that are enriched for DIND outliers, for each population.

GWAS diseases categories	Population	Risk[b]	Protection[c]	NR[d]	Genes	Position	References
Cancers							
Breast cancer	AFR		rs4415084		Intergenic	5	(Fletcher et al. 2011)
	EUR	rs11249433			Intergenic	1	(Thomas et al. 2009)
	ASN	rs2180341			*ECHDC1, RNF146*	6q22.33	(Gold et al. 2008)
Esophageal cancer	ASN	rs1229984[e]			*ADH6, ADH1B*	4q23	(McKay et al. 2011)
Leukemia	AFR		rs4795519		*WSB1, FAM27L*	17q11.1-q11.2	(Kim, Lee, et al. 2011)
Pancreatic cancer	AFR		rs372883		*BACH1*	21q22.11	(Wu et al. 2012)
Primary biliary cirrhosis	AFR		rs9303277		*IKZF3, ZPBP2, GSDMB, ORMDL3*	17q12-q21	(Liu et al. 2010)
Prostate cancer	EUR			rs7584330[f]	Intergenic	2	(Kote-Jarai et al. 2011)
		rs2121875			*FGF10*	5p13-p12	(Kote-Jarai et al. 2011)
			rs10875943		*PRPH*	12q12-q13	(Kote-Jarai et al. 2011)
				rs2292884[f]	*MLPH*	2q37.3	(Schumacher et al. 2011)
	ASN	rs2121875			*FGF10*	5p13-p12	(Kote-Jarai et al. 2011)
			rs17181170		*NR*	3	(Eeles et al. 2009)
Testicular cancer	EUR	rs4474514[m]			*KITLG*	12q22	(Kanetsky et al. 2009)
		rs995030[m]			*KITLG*	12q22	(Rapley et al. 2009)

				rs3782181[m]	*KITLG*	12q22	(Turnbull et al. 2010)
	ASN	rs4474514[n]			*KITLG*	12q22	(Kanetsky et al. 2009)
		rs995030[n]			*KITLG*	12q22	(Rapley et al. 2009)
				rs3782181[n]	*KITLG*	12q22	(Turnbull et al. 2010)
Upper aerodigestive tract cancers	EUR		rs4767364		*ALDH2*	12q24.2	(McKay et al. 2011)
	ASN	rs1229984[e]			*ADH1B*	4q23	(McKay et al. 2011)
Autoimmune diseases							
Amyotrophic lateral sclerosis	EUR	rs1541160			*KIFAP3*	1q24.2	(Landers et al. 2009)
Asthma	EUR	rs1588265			*PDE4D*	5q12	(Himes et al. 2009)
Celiac disease	EUR	rs653178[q]			*SH2B3, ATXN2*	12q24-q24.1	(Zhernakova et al. 2011)
	ASN	rs917997[g]			*IL18RAP, IL18R1, IL1RL1, IL1RL2*	2q12	(Dubois et al. 2010)
Chronic kidney disease	EUR		rs653178[q]		*ATXN2*	12q24.1	(Kottgen et al. 2010)
		rs267734			*ANXA9,FAM63A, PRUNE, BNIPL,LASS2,SETDB1*	1q21-q21.3	(Kottgen et al. 2010)
				rs2453533[p]	*GATM, SPATA5L1*	15q21.1	(Kottgen et al. 2010)
			rs2467853[p]		*GATM, SPATA5L1*	15q21.1	(Kottgen et al. 2010)
Crohn's disease	AFR	rs3197999[h]			*MST1, GPX1,*	3p21-p21.31	(Barrett et al. 2008)

				BSN		
		rs9858542[h]		*MST1,BSN*	3p21-p21.31	(Barrett et al. 2008)
			rs2301436	*CCR6, FGFR10P*	6q27	(McGovern, Jones, et al. 2010)
	EUR	rs2058660		*IL18RAP, IL12RL2, IL18R1, IL1RL1*	2q12	(Franke et al. 2010)
		rs3091338[i]		*IL3, ACSL6, P4HA2, PDLIM4, SLC22A4*	5q31-q31.1	(Kenny et al. 2012)
		rs12521868[i]		*SLC22A4, SLC22A5, IRF1, IL3*	5q31.1	(Franke et al. 2010)
		rs2188962		Intergenic	5	(McGovern, Jones, et al. 2010)
	ASN	rs2058660[g]		*IL18RAP, IL12RL2, IL18R1, IL1RL1*	2q12	(Franke et al. 2010)
		rs9286879		Intergenic	1	(Barrett et al. 2008)
Inflammatory bowel disease	EUR	rs9271366[j]		*MHC*	6p21.3	(Okada et al. 2011)
Multiple sclerosis	AFR		rs11581062	*SLC30A7*	1p21.2	(Sawcer et al. 2011)
	EUR	rs3135388[j]		*HLA-DRB1, HLA-DRA*	6p21.3	(De Jager et al. 2009)
		rs9271366[j]		*HLA-DRB1*	6p21.3	(Nischwitz et al. 2010)
			rs180515	*RPS6KB1*	17q23.1	(Sawcer et al. 2011)
		rs3129934[j]		*HLA-DRB1, C6orf10*	6p21.3	(Martinelli-Boneschi et al. 2012)
Psoriasis	ASN		rs10782001	*FBXL19, POL3S*	16p11.2	(Stuart et al. 2010)

Rheumatoid arthritis	EUR	rs6910071			*HLA-DRB1*	6p21.3	(Stahl et al. 2010)
		rs653178[k]			*SH2B3*	12q24	(Zhernakova et al. 2011)
			rs3761847		*TRAF1-C5*	9q33-q34	(Stahl et al. 2010)
Systemic lupus erythematosus	ASN	rs7197475			*NR*	16	(Han et al. 2009)
Type 1 diabetes	EUR	rs17696736[k]			*SH2B3, LNK, TRAFD1, PTPN11, C12orf30*	12q-q24-q24.13	(Cooper et al. 2008)
		rs3184504[k]			*SH2B3*	12q24	(Barrett, Clayton, et al. 2009)
Ulcerative colitis	AFR	rs3197999[h,o]			*MST1*	3p21	(McGovern, Gardet, et al. 2010)
				rs9858542[o]	*MST1*	3p21	(Barrett, Lee, et al. 2009)
Other diseases							
Age-related macular degeneration	EUR	rs1061170[l]			*CFH*	1q32	(Yu et al. 2011)
		rs1329424[l]			*CFH*	1q32	(Chen et al. 2010)
				rs1999930	*FRK, COL10A1*	6q21-q22.3	(Yu et al. 2011)
Alcoholism	AFR		rs6943555		*AUTS2*	7q11.22	(Schumann et al. 2011)
	ASN		rs6943555		*AUTS2*	7q11.22	(Schumann et al. 2011)
Coronary heart disease	EUR		rs17609940		*ANKS1A*	6p21.31	(Schunkert et al. 2011)
Dupuytren's disease	AFR			rs611744	*EIF3E, RSPO2*	8q22-q23	(Dolmans et al. 2011)
Glaucoma	EUR		rs10483727		*SIX1, SIX6*	14q23.1	(Osman et al. 2012)

Hirschsprung's disease	ASN	rs16879552			*NRG1*	8p12	(Garcia-Barcelo et al. 2009)
Myopia (pathological)	ASN			rs10034228	*MYP11*	4q22-q27	(Li et al. 2011)
Permanent tooth development	EUR	rs7924176			*ADK*	10q11-q24	(Geller et al. 2011)
Polycystic ovary syndrome	EUR		rs2479106		*DENND1A*	9q33.3	(Chen et al. 2011)
	ASN		rs2479106		*DENND1A*	9q33.3	(Chen et al. 2011)
Progressive supranuclear palsy	EUR			rs242557	*MAPT*	17q21.1	(Hoglinger et al. 2011)
				rs7571971	*EIF2AK3*	2p12	(Hoglinger et al. 2011)
	ASN			rs7571971	*EIF2AK3*	2p12	(Hoglinger et al. 2011)
Psychiatric diseases	AFR	rs7914558			*CNNM2*	10q24.32	(Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium 2011)
		rs1625579			*MIR137*	1p21.3	
	ASN	rs1625579			*MIR137*	1p21.3	
Type 2 diabetes	AFR		rs7178572		*HMG20A*	15q24	(Kooner et al. 2011)
	EUR		rs849134		*JAZF1*	7p15.2-p15.1	(Voight et al. 2010)

[a]The list of diseases enriched in DIND outliers can be found on supplementary table S14, Supplementary Material online.
[b]The selected allele (derived allele) is associated to an increase in the risk to develop the disease.
[c]The selected allele (derived allele) is not the risk allele as defined in NHGRI GWAS database.
[d]The risk-allele wasn't reported on the NHGRI GWAS database.
[e,f,g,h,i,j,k,l]These SNPs are in linkage disequilibrium and show the same direction of selection. Each set of SNPs in LD is counted as one SNP.
[m,n,o,p]These SNPs are in linkage disequilibrium and at least one of each set has non-reported risk allele. Each set of SNPs is counted as one SNP and NR-risk allele are excluded from the total.
[q]These SNPs are in linkage disequilibrium , and present different direction of selection for two different autoimmune diseases. We have removed them from the total.

Table S20. Number of selected alleles associated to cholesterol level (HDL, LDL or total) in all the populations.

Population	Cholesterol type[a]	Increase[b]	Decrease[c]	Genes	Position	References
AFR	HDL	rs16942887		*LCAT*	16q22.1	(Teslovich et al. 2010)
	HDL		rs7134594[d]	*MMAB,MVK*	12q24	(Teslovich et al. 2010)
	HDL	rs4759375[d]		*SBNO1*	12q24.31	(Teslovich et al. 2010)
Total AFR		**1**	**0**			
EUR	Total	rs2338104[e]		*MMAB,MVK*	12q24	(Kathiresan et al. 2009)
	Total		rs7570971	*RAB3GAP1*	2q21.3	(Teslovich et al. 2010)
	LDL	rs12916[f]		*HMGCR*	5q13.3-q14	(Waterworth et al. 2010)
	LDL		rs11065987[e]	*BRAP*	12q24	(Teslovich et al. 2010)
	LDL		rs7703051[f]	*HMGCR*	5q13.3-q14	(Burkhardt et al. 2008)
	LDL	rs12654264[f]		*HMGCR*	5q13.3-q14	(Kim, Go, et al. 2011)
	LDL	rs3846663[f]		*HMGCR*	5q13.3-q14	(Kathiresan et al. 2009)
	HDL	rs7134594[e]		*MMAB,MVK*	12q24	(Teslovich et al. 2010)
Total EUR		**0**	**1**			
ASN	HDL	rs3136441		*LRP4,NR1H3*	11p11.2	(Teslovich et al. 2010)
	HDL		rs7134594	*MMAB,MVK*	12q24	(Teslovich et al. 2010)
	LDL	rs7206971		*OBSPL7*	17q21.32	(Teslovich et al. 2010)
Total ASN		**2**	**1**			

[a]LDL means low-density lipoprotein, HDL means high-density lipoprotein, and Total cholesterol is the sum of HDL and LDL levels.
[b]The selected allele (derived allele) is associated to high cholesterol levels.
[c]The selected allele (derived allele) is associated to low cholesterol level.
[d,e,f]These SNPs are in linkage disequilibrium, and present different direction of selection. We have removed them from the total.

References

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061-1073.

Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, et al. 2009. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet* 41:703-707.

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, et al. 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* 40:955-962.

Barrett JC, Lee JC, Lees CW, Prescott NJ, Anderson CA, Phillips A, Wesley E, Parnell K, Zhang H, Drummond H, et al. 2009. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat Genet* 41:1330-1334.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025-2035.

Burkhardt R, Kenny EE, Lowe JK, Birkeland A, Josowitz R, Noel M, Salit J, Maller JB, Pe'er I, Daly MJ, et al. 2008. Common SNPs in HMGCR in micronesians and whites associated with LDL-cholesterol levels affect alternative splicing of exon13. *Arterioscler Thromb Vasc Biol* 28:2078-2084.

Carty CL, Johnson NA, Hutter CM, Reiner AP, Peters U, Tang H, Kooperberg C. 2012. Genome-wide association study of body height in African Americans: the Women's Health Initiative SNP Health Association Resource (SHARe). *Hum Mol Genet* 21:711-720.

Chen W, Stambolian D, Edwards AO, Branham KE, Othman M, Jakobsdottir J, Tosakulwong N, Pericak-Vance MA, Campochiaro PA, Klein ML, et al. 2010. Genetic variants near TIMP3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration. *Proc Natl Acad Sci U S A* 107:7401-7406.

Chen ZJ, Zhao H, He L, Shi Y, Qin Y, Shi Y, Li Z, You L, Zhao J, Liu J, et al. 2011. Genome-wide association study identifies susceptibility loci for polycystic ovary syndrome on chromosome 2p16.3, 2p21 and 9q33.3. *Nat Genet* 43:55-59.

Cooper JD, Smyth DJ, Smiles AM, Plagnol V, Walker NM, Allen JE, Downes K, Barrett JC, Healy BC, Mychaleckyj JC, et al. 2008. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat Genet* 40:1399-1401.

De Jager PL, Jia X, Wang J, de Bakker PI, Ottoboni L, Aggarwal NT, Piccio L, Raychaudhuri S, Tran D, Aubin C, et al. 2009. Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat Genet* 41:776-782.

Dolmans GH, Werker PM, Hennies HC, Furniss D, Festen EA, Franke L, Becker K, van der Vlies P, Wolffenbuttel BH, Tinschert S, et al. 2011. Wnt signaling and Dupuytren's disease. *N Engl J Med* 365:307-317.

Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, Zhernakova A, Heap GA, Adany R, Aromaa A, et al. 2010. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* 42:295-302.

Eeles RA, Kote-Jarai Z, Al Olama AA, Giles GG, Guy M, Severi G, Muir K, Hopper JL, Henderson BE, Haiman CA, et al. 2009. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat Genet* 41:1116-1121.

Elks CE, Perry JR, Sulem P, Chasman DI, Franceschini N, He C, Lunetta KL, Visser JA, Byrne EM, Cousminer DL, et al. 2010. Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. *Nat Genet* 42:1077-1085.

Fletcher O, Johnson N, Orr N, Hosking FJ, Gibson LJ, Walker K, Zelenika D, Gut I, Heath S, Palles C, et al. 2011. Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. *J Natl Cancer Inst* 103:425-435.

Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, et al. 2010. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 42:1118-1125.
Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851-861.
Garcia-Barcelo MM, Tang CS, Ngan ES, Lui VC, Chen Y, So MT, Leon TY, Miao XP, Shum CK, Liu FQ, et al. 2009. Genome-wide association study identifies NRG1 as a susceptibility locus for Hirschsprung's disease. *Proc Natl Acad Sci U S A* 106:2694-2699.
Geller F, Feenstra B, Zhang H, Shaffer JR, Hansen T, Esserlind AL, Boyd HA, Nohr EA, Timpson NJ, Fatemifar G, et al. 2011. Genome-wide association study identifies four loci associated with eruption of permanent teeth. *PLoS Genet* 7:e1002275.
Gold B, Kirchhoff T, Stefanov S, Lautenberger J, Viale A, Garber J, Friedman E, Narod S, Olshen AB, Gregersen P, et al. 2008. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. *Proc Natl Acad Sci U S A* 105:4340-4345.
Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD. 2011. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A* 108:11983-11988.
Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, Zusmanovich P, Sulem P, Thorlacius S, Gylfason A, Steinberg S, et al. 2008. Many sequence variants affecting diversity of adult human height. *Nat Genet* 40:609-615.
Han JW, Zheng HF, Cui Y, Sun LD, Ye DQ, Hu Z, Xu JH, Cai ZM, Huang W, Zhao GP, et al. 2009. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat Genet* 41:1234-1237.
Himes BE, Hunninghake GM, Baurley JW, Rafaels NM, Sleiman P, Strachan DP, Wilk JB, Willis-Owen SA, Klanderman B, Lasky-Su J, et al. 2009. Genome-wide association analysis identifies PDE4D as an asthma-susceptibility gene. *Am J Hum Genet* 84:581-593.
Hoglinger GU, Melhem NM, Dickson DW, Sleiman PM, Wang LS, Klei L, Rademakers R, de Silva R, Litvan I, Riley DE, et al. 2011. Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. *Nat Genet* 43:699-705.
Johansson A, Marroni F, Hayward C, Franklin CS, Kirichenko AV, Jonasson I, Hicks AA, Vitart V, Isaacs A, Axenovich T, et al. 2009. Common variants in the JAZF1 gene associated with height identified by linkage and genome-wide association analysis. *Hum Mol Genet* 18:373-380.
Kanetsky PA, Mitra N, Vardhanabhuti S, Li M, Vaughn DJ, Letrero R, Ciosek SL, Doody DR, Smith LM, Weaver J, et al. 2009. Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. *Nat Genet* 41:811-815.
Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, Kaplan L, Bennett D, Li Y, Tanaka T, et al. 2009. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet* 41:56-65.
Kayser M, Liu F, Janssens AC, Rivadeneira F, Lao O, van Duijn K, Vermeulen M, Arp P, Jhamai MM, van Ijcken WF, et al. 2008. Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. *Am J Hum Genet* 82:411-423.
Kenny EE, Pe'er I, Karban A, Ozelius L, Mitchell AA, Ng SM, Erazo M, Ostrer H, Abraham C, Abreu MT, et al. 2012. A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. *PLoS Genet* 8:e1002559.
Kim DH, Lee ST, Won HH, Kim S, Kim MJ, Kim HJ, Kim SH, Kim JW, Kim HJ, Kim YK, et al. 2011. A genome-wide association study identifies novel loci associated with susceptibility to chronic myeloid leukemia. *Blood* 117:6906-6911.

Kim YJ, Go MJ, Hu C, Hong CB, Kim YK, Lee JY, Hwang JY, Oh JH, Kim DJ, Kim NH, et al. 2011. Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. *Nat Genet* 43:990-995.

Kooner JS, Saleheen D, Sim X, Sehmi J, Zhang W, Frossard P, Been LF, Chia KS, Dimas AS, Hassanali N, et al. 2011. Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet* 43:984-989.

Kote-Jarai Z, Olama AA, Giles GG, Severi G, Schleutker J, Weischer M, Campa D, Riboli E, Key T, Gronberg H, et al. 2011. Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. *Nat Genet* 43:785-791.

Kottgen A, Pattaro C, Boger CA, Fuchsberger C, Olden M, Glazer NL, Parsa A, Gao X, Yang Q, Smith AV, et al. 2010. New loci associated with kidney function and chronic kidney disease. *Nat Genet* 42:376-384.

Landers JE, Melki J, Meininger V, Glass JD, van den Berg LH, van Es MA, Sapp PC, van Vught PW, McKenna-Yasek DM, Blauw HM, et al. 2009. Reduced expression of the Kinesin-Associated Protein 3 (KIFAP3) gene increases survival in sporadic amyotrophic lateral sclerosis. *Proc Natl Acad Sci U S A* 106:9004-9009.

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, et al. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467:832-838.

Laval G, Excoffier L. 2004. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics* 20:2485-2487.

Laval G, Patin E, Barreiro LB, Quintana-Murci L. 2010. Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS One* 5:e10284.

Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, Eyheramendy S, Voight BF, Butler JL, Guiducci C, et al. 2008. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet* 40:584-591.

Li Z, Qu J, Xu X, Zhou X, Zou H, Wang N, Li T, Hu X, Zhao Q, Chen P, et al. 2011. A genome-wide association study reveals association between common variants in an intergenic region of 4q25 and high-grade myopia in the Chinese Han population. *Hum Mol Genet* 20:2861-2868.

Liu X, Invernizzi P, Lu Y, Kosoy R, Lu Y, Bianchi I, Podda M, Xu C, Xie G, Macciardi F, et al. 2010. Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis. *Nat Genet* 42:658-660.

Martinelli-Boneschi F, Esposito F, Brambilla P, Lindstrom E, Lavorgna G, Stankovich J, Rodegher M, Capra R, Ghezzi A, Coniglio G, et al. 2012. A genome-wide association study in progressive multiple sclerosis. *Mult Scler* 18:1384-1394.

McGovern DP, Gardet A, Torkvist L, Goyette P, Essers J, Taylor KD, Neale BM, Ong RT, Lagace C, Li C, et al. 2010. Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nat Genet* 42:332-337.

McGovern DP, Jones MR, Taylor KD, Marciante K, Yan X, Dubinsky M, Ippoliti A, Vasiliauskas E, Berel D, Derkowski C, et al. 2010. Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. *Hum Mol Genet* 19:3468-3476.

McKay JD, Truong T, Gaborieau V, Chabrier A, Chuang SC, Byrnes G, Zaridze D, Shangina O, Szeszenia-Dabrowska N, Lissowska J, et al. 2011. A genome-wide association study of upper aerodigestive tract cancers conducted within the INHANCE consortium. *PLoS Genet* 7:e1001333.

N'Diaye A, Chen GK, Palmer CD, Ge B, Tayo B, Mathias RA, Ding J, Nalls MA, Adeyemo A, Adoue V, et al. 2011. Identification, replication, and fine-mapping of Loci associated with adult height in individuals of african ancestry. *PLoS Genet* 7:e1002298.

Nischwitz S, Cepok S, Kroner A, Wolf C, Knop M, Muller-Sarnowski F, Pfister H, Roeske D, Rieckmann P, Hemmer B, et al. 2010. Evidence for VAV2 and ZNF433 as susceptibility genes for multiple sclerosis. *J Neuroimmunol* 227:162-166.

Okada Y, Yamazaki K, Umeno J, Takahashi A, Kumasaka N, Ashikawa K, Aoi T, Takazoe M, Matsui T, Hirano A, et al. 2011. HLA-Cw*1202-B*5201-DRB1*1502 haplotype increases risk for ulcerative colitis but reduces risk for Crohn's disease. *Gastroenterology* 141:864-871 e861-865.

Osman W, Low SK, Takahashi A, Kubo M, Nakamura Y. 2012. A genome-wide association study in the Japanese population confirms 9p21 and 14q23 as susceptibility loci for primary open angle glaucoma. *Hum Mol Genet* 21:2836-2842.

Rapley EA, Turnbull C, Al Olama AA, Dermitzakis ET, Linger R, Huddart RA, Renwick A, Hughes D, Hines S, Seal S, et al. 2009. A genome-wide association study of testicular germ cell tumor. *Nat Genet* 41:807-810.

Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, Moutsianas L, Dilthey A, Su Z, Freeman C, Hunt SE, et al. 2011. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* 476:214-219.

Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. 2011. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* 43:969-976.

Schumacher FR, Berndt SI, Siddiq A, Jacobs KB, Wang Z, Lindstrom S, Stevens VL, Chen C, Mondul AM, Travis RC, et al. 2011. Genome-wide association study identifies new prostate cancer susceptibility loci. *Hum Mol Genet* 20:3867-3875.

Schumann G, Coin LJ, Lourdusamy A, Charoen P, Berger KH, Stacey D, Desrivieres S, Aliev FA, Khan AA, Amin N, et al. 2011. Genome-wide association and genetic functional studies identify autism susceptibility candidate 2 gene (AUTS2) in the regulation of alcohol consumption. *Proc Natl Acad Sci U S A* 108:7119-7124.

Schunkert H, Konig IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, Preuss M, Stewart AF, Barbalic M, Gieger C, et al. 2011. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet* 43:333-338.

Soranzo N, Rivadeneira F, Chinappen-Horsley U, Malkina I, Richards JB, Hammond N, Stolk L, Nica A, Inouye M, Hofman A, et al. 2009. Meta-analysis of genome-wide scans for human adult stature identifies novel Loci and associations with measures of skeletal frame size. *PLoS Genet* 5:e1000445.

Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, Li Y, Kurreeman FA, Zhernakova A, Hinks A, et al. 2010. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet* 42:508-514.

Stuart PE, Nair RP, Ellinghaus E, Ding J, Tejasvi T, Gudjonsson JE, Li Y, Weidinger S, Eberlein B, Gieger C, et al. 2010. Genome-wide association analysis identifies three psoriasis susceptibility loci. *Nat Genet* 42:1000-1004.

Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G, et al. 2007. Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet* 39:1443-1452.

Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP, Ripatti S, Chasman DI, Willer CJ, et al. 2010. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466:707-713.

Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE, Hutchinson A, Wang Z, Yu K, et al. 2009. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat Genet* 41:579-584.

Turnbull C, Rapley EA, Seal S, Pernet D, Renwick A, Hughes D, Ricketts M, Linger R, Nsengimana J, Deloukas P, et al. 2010. Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer. *Nat Genet* 42:604-607.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci U S A* 102:18508-18513.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol* 4:e72.

Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, et al. 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 42:579-589.

Waterworth DM, Ricketts SL, Song K, Chen L, Zhao JH, Ripatti S, Aulchenko YS, Zhang W, Yuan X, Lim N, et al. 2010. Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arterioscler Thromb Vasc Biol* 30:2264-2276.

Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, Freathy RM, Perry JR, Stevens S, Hall AS, et al. 2008. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* 40:575-583.

Wu C, Miao X, Huang L, Che X, Jiang G, Yu D, Yang X, Cao G, Hu Z, Zhou Y, et al. 2012. Genome-wide association study identifies five loci associated with susceptibility to pancreatic cancer in Chinese populations. *Nat Genet* 44:62-66.

Yu Y, Bhangale TR, Fagerness J, Ripke S, Thorleifsson G, Tan PL, Souied EH, Richardson AJ, Merriam JE, Buitendijk GH, et al. 2011. Common variants near FRK/COL10A1 and VEGFA are associated with advanced age-related macular degeneration. *Hum Mol Genet* 20:3699-3709.

Zhernakova A, Stahl EA, Trynka G, Raychaudhuri S, Festen EA, Franke L, Westra HJ, Fehrmann RS, Kurreeman FA, Thomson B, et al. 2011. Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet* 7:e1002004.

Annexe B

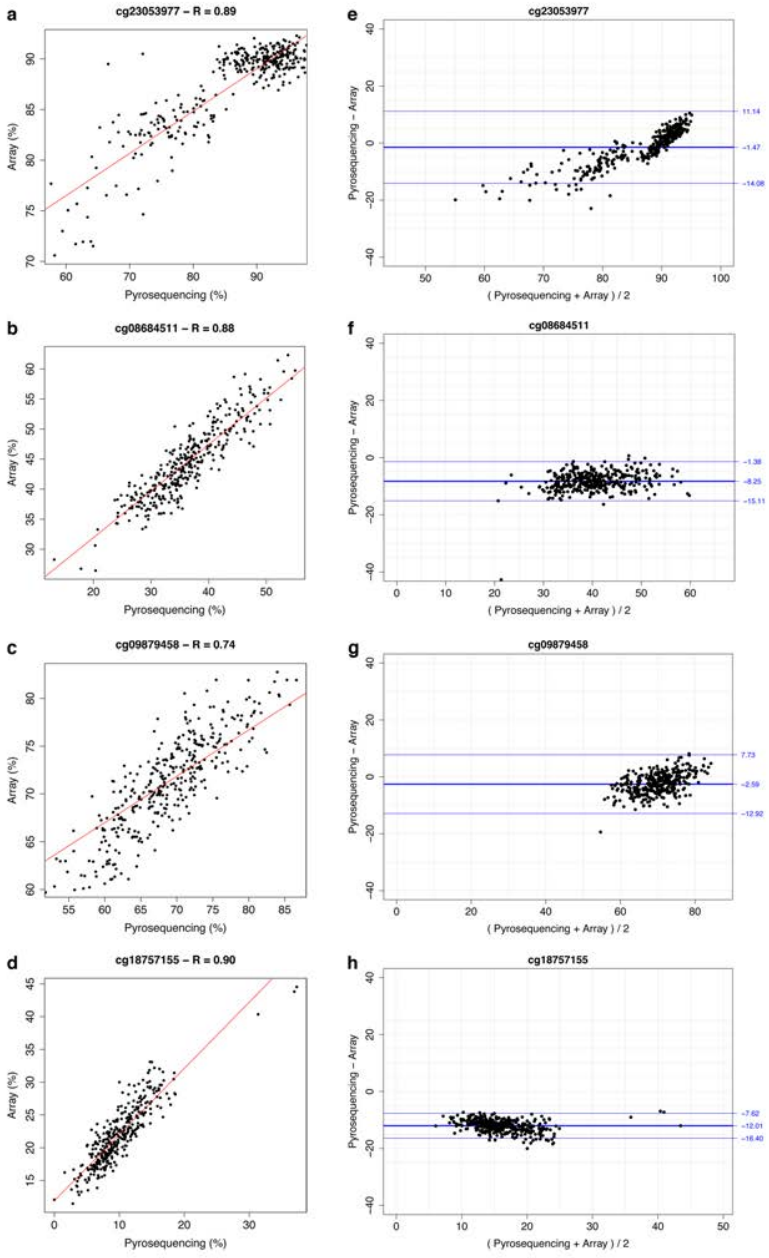
Compléments d'informations pour l'article 2

Supplementary Information for:

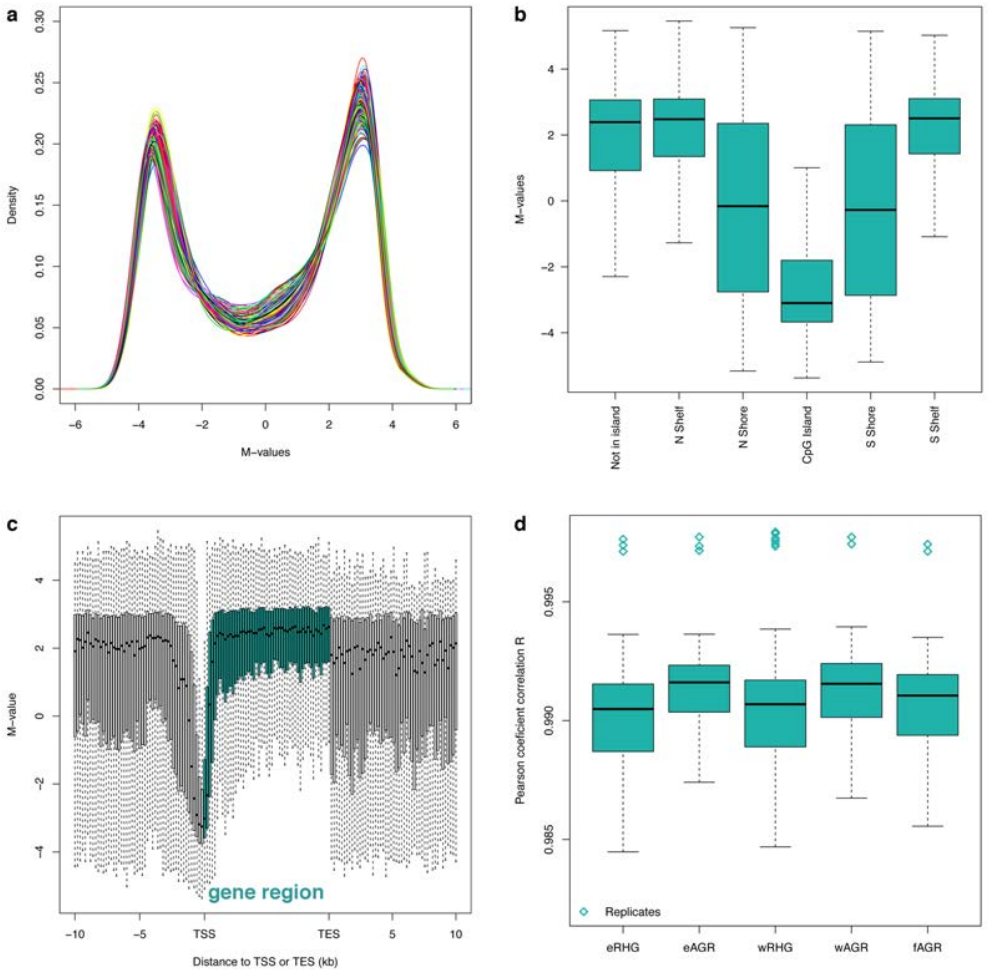
The Epigenomic Landscape of African Rainforest Hunter-Gatherers and Farmers

Maud Fagny, Etienne Patin, Julia L. MacIsaac, Maxime Rotival, Timothée Flutre, Meaghan J. Jones, Katherine J. Siddle, Hélène Quach, Christine Harmant, Lisa M. McEwen, Alain Froment, Evelyne Heyer, Antoine Gessain, Edouard Betsem, Patrick Mouguiama-Daouda, Jean-Marie Hombert, George H. Perry, Luis B. Barreiro, Michael S. Kobor, Lluis Quintana-Murci

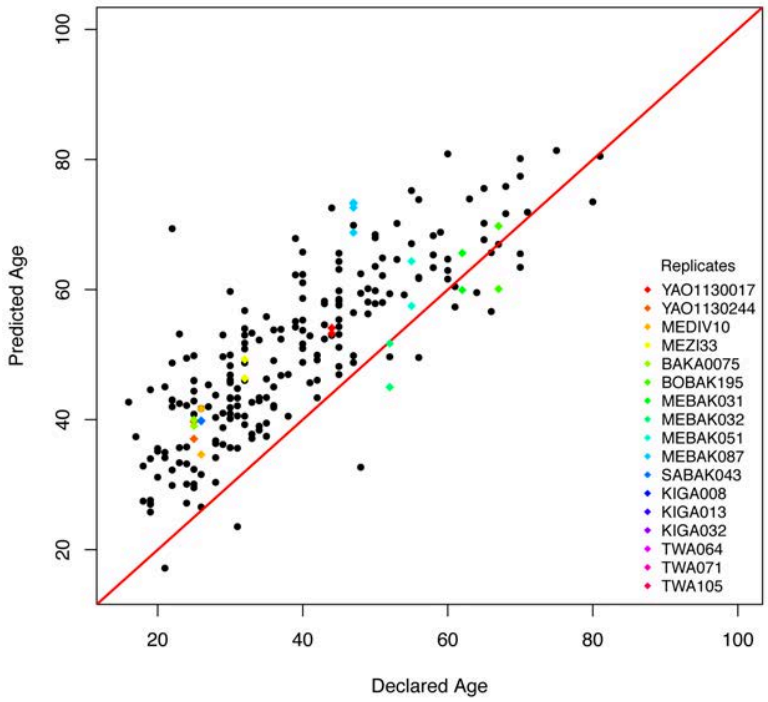
This PDF file includes:



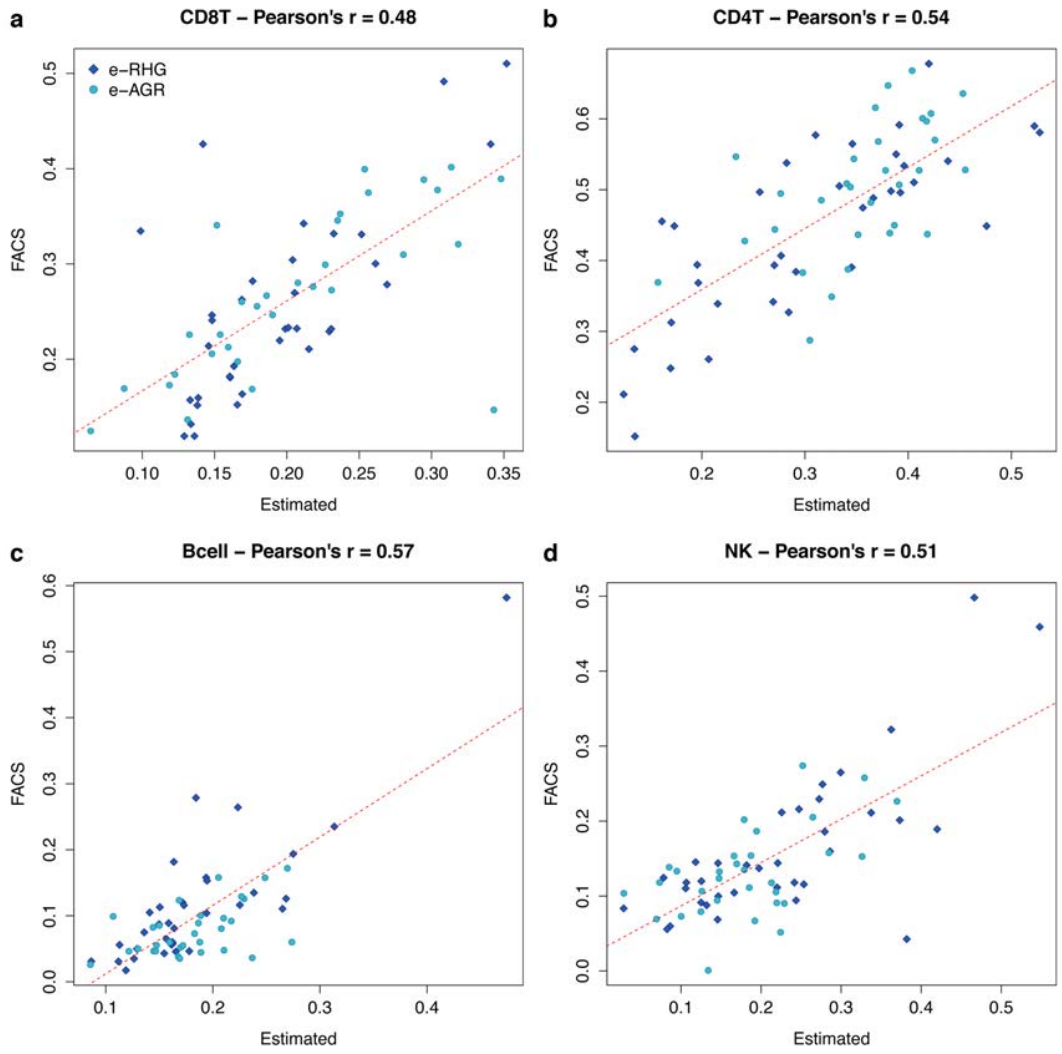
Supplementary Fig. 1. Comparison of DNA methylation profiles obtained by array and pyrosequencing. (**a-d**) Proportion of methylation obtained by pyrosequencing and array. Red lines correspond to the fit obtained by linear regression (Array ~ Pyrosequencing). Pearson's *R* coefficients are indicated on top of each panel. (**e-h**) Bland-Altman plot comparing pyrosequencing and array methods. Thick blue lines indicate the mean of Pyrosequencing – Array values. Thin blue lines represent limits of the 95% confidence interval. (**a,e**) cg23053977 in the 6p12.3 region (enhancer); (**b,f**) cg08684511 in the gene *COL23A1*; (**c,g**) cg09879458 in the gene *RORA*; and (**d,h**) cg18757155 in the gene *ADAM28*.



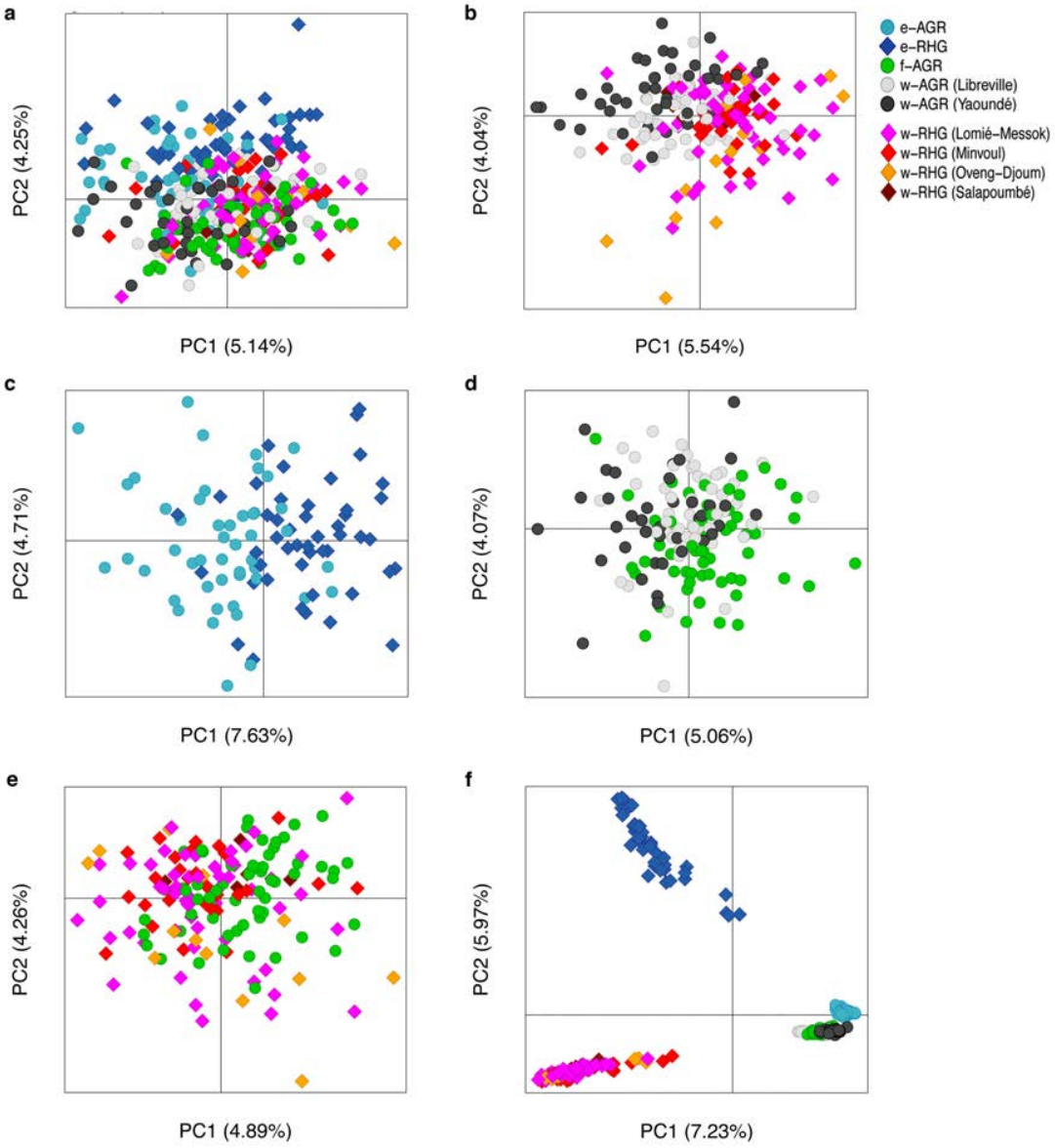
Supplementary Fig. 2. General characteristics of DNA methylation profiles. (**a**) Distribution of M-values after filtering and normalisation. Each line represents the distribution of all M-values for a sample. Negative values correspond to unmethylated sites, null values correspond to hemi-methylated sites and positive values correspond to fully methylated sites. (**b**) CpG island-related distribution of M-values for a representative sample (NZE7003). "Shore" refers to the 0-2kb region from the CpG island on each side; "Shelf" refers to the 2-4kb region from the CpG island on each side; "N" refers to the shore or shelf in 5' of the CpG island. "S" refers to the shore or shelf in 3' of the CpG island. All samples showed the same trends (data not shown). (**c**) Distribution of M-values within 10 kb around gene regions. Sites outside of gene regions were binned in 200bp bins from the TSS or the TES. Gene regions, in blue, were split into 50 equally-sized bins, regardless of the length of the gene. (**d**) Distribution of Pearson's correlation coefficients between methylation M-values of all pairs of samples in each population. Diamonds indicates the very high correlation coefficients between technical replicates (Pearson $R > 0.997$).



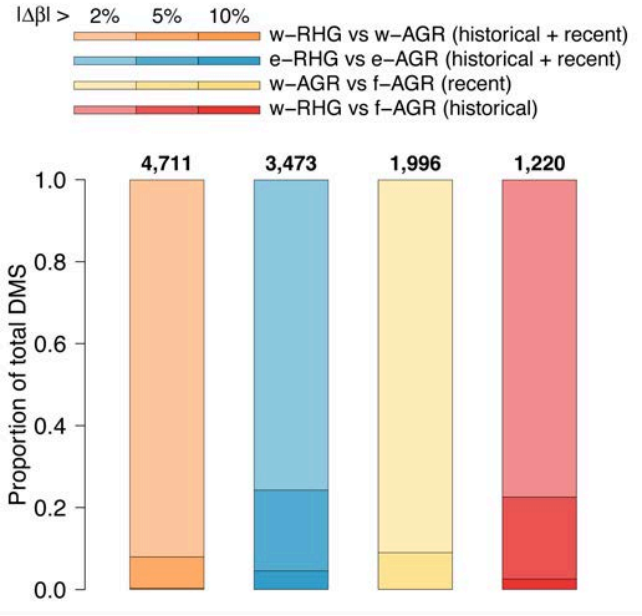
Supplementary Fig. 3. Correlation between declared and predicted ages. Plot showing the correlation between reported ages and predicted ages, the latter imputed from methylation data for all samples using an elastic net regression model. The colored diamonds represent technical replicates. The slight variance of predicted ages for the same reported age could be due to a general lack of accuracy of the declared age, owing to the absence of systematic registration of people at birth.



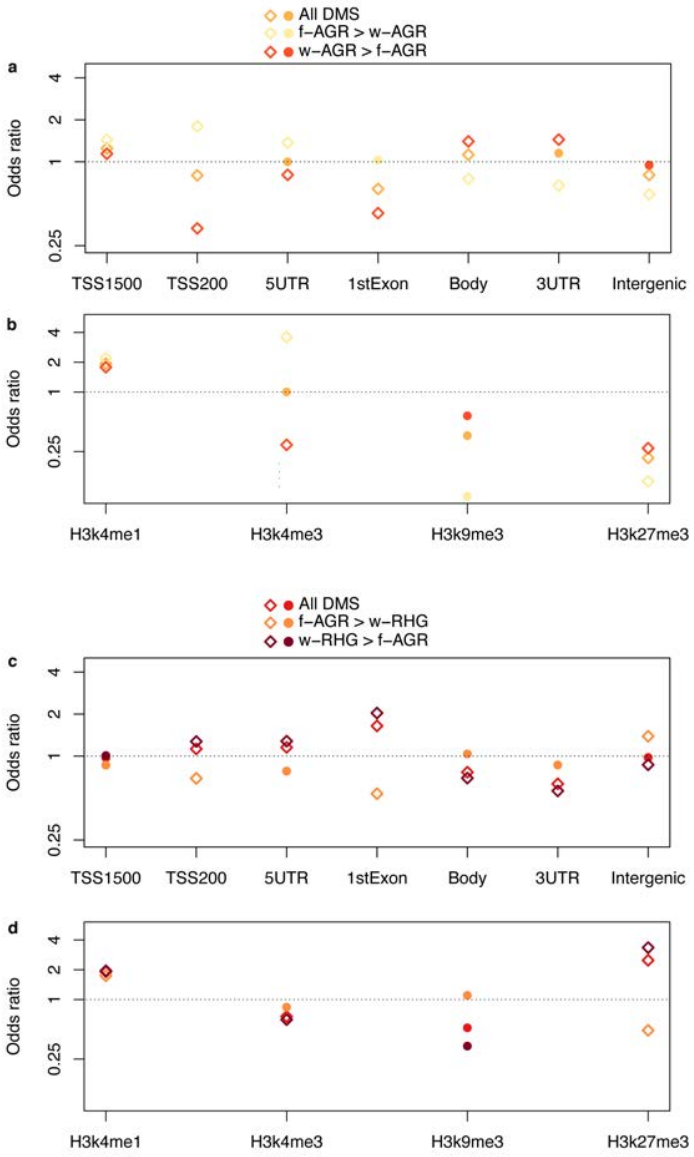
Supplementary Fig. 4. Correlation between predicted and measured proportions of different cell types. Plots showing the correlation between estimated cell proportions based on the DNA methylation signature of each of the principal immune cell components and observed cell proportions based on FACS data for (**a**) $CD8^+$ T cells, (**b**) $CD4^+$ T cells, (**c**) B cells and (**d**) NK cells. Dark blue diamonds represent e-RHG samples and light blue circles represent e-AGR samples. Red lines, in each plot, indicate the fitted model of a linear regression for FACS-based ~ estimated cell counts.



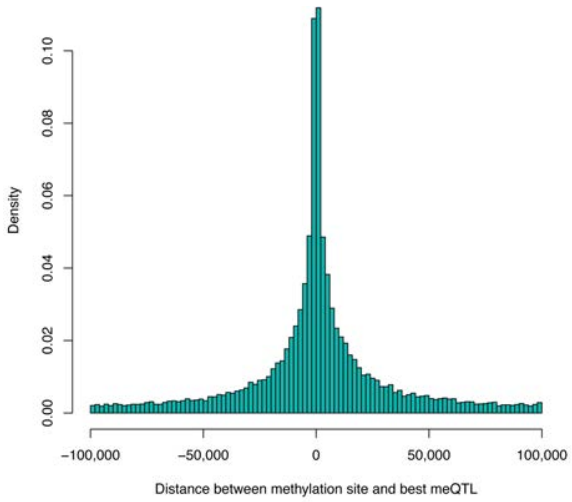
Supplementary Fig. 5. Principal component analysis (PCA) of global DNA methylation data and genotyping data. (**a**) PCA of global methylation patterns including all population of our study (the two populations of w-AGR and the four populations of w-RHG are labeled in different colors to show the great similarity in methylation patterns within AGR and RHG samples). (**b**) PCA of genome-wide methylation profiles for w-RHG and w-AGR samples. (**c**) PCA of methylation profiles for e-RHG and e-AGR. (**d**) PCA of methylation profiles for f-AGR and w-AGR samples. (**e**) PCA of methylation profiles for w-RHG and f-AGR samples. (**f**) PCA of the genotype data for study populations, based on 456,507 independent SNPs genome-wide (the two populations of w-AGR and the four populations of w-RHG are labeled in different colors to show the great similarity in genotype patterns within AGR and RHG samples). The proportion of the variance explained by PC1 and PC2 is indicated on the axes.



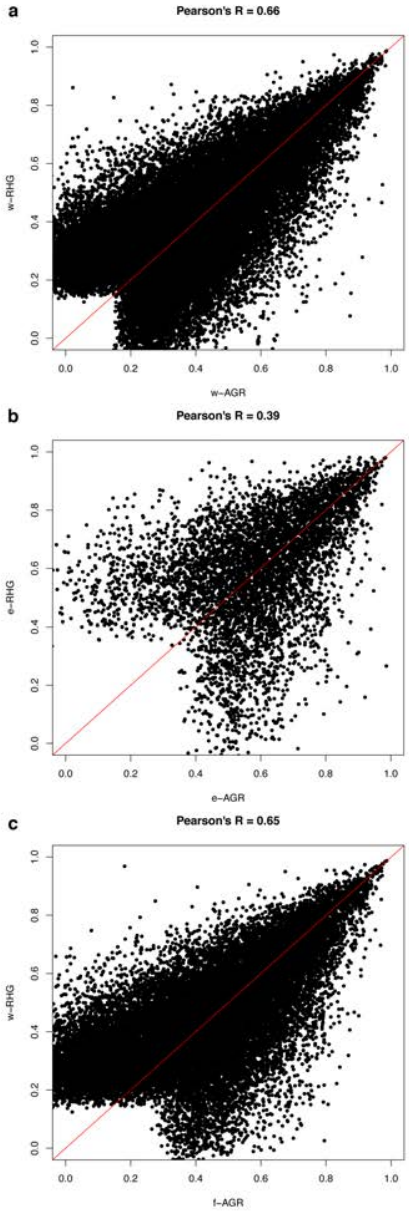
Supplementary Fig. 6. Amplitude of methylation differences. Proportions of DMS that explain mean methylation levels higher than 5% and 10% among DMS presenting a $\Delta\beta > 2\%$, for each population comparison. Numbers on the top of the bars represent the number of DMS in each category. Historical DMS presented a significant enrichment in DMS with mean mehtylation levels higher than 5% with respect to recent DMS ($P = 10^{-16}$, χ^2-test). Similarly, historical DMS presented 32 DMS with mean methylation values higher than 10%, whereas recent DMS had none.



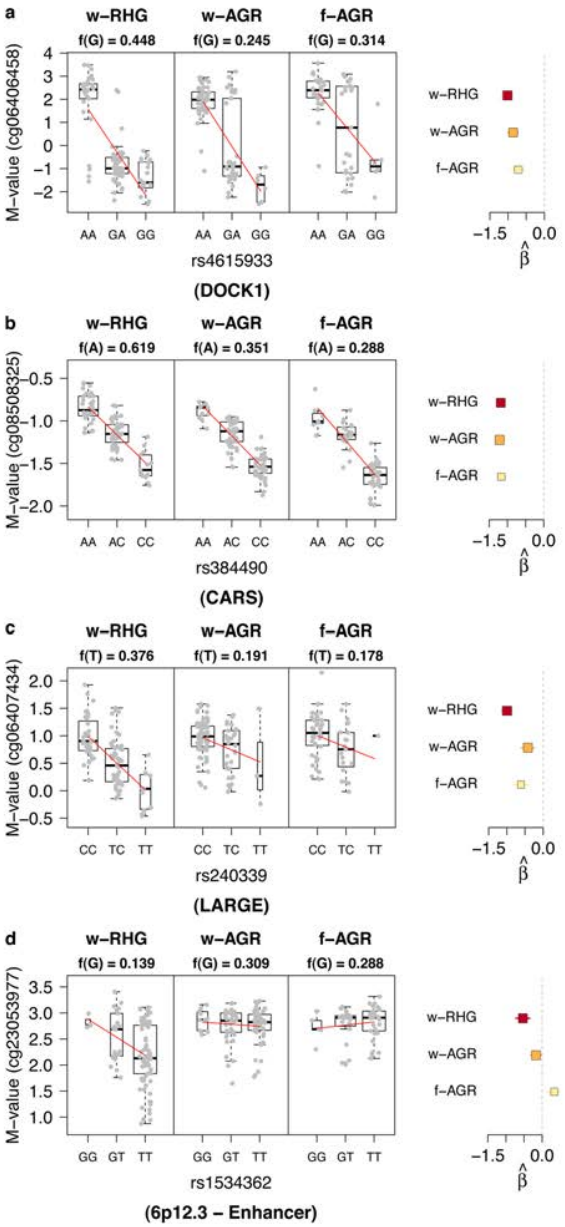
Supplementary Fig. 7. Functional features of DMS. (**a**) Odds ratios measuring the enrichment of different genomic locations in recent DMS (5,765 DMS) for all DMS (in orange), DMS more highly methylated in f-AGR with respect to w-AGR (f-AGR > w-AGR, in yellow) and DMS more highly methylated in w-AGR with respect to f-AGR (w-AGR > f-AGR, in red). (**b**) Odds ratios measuring the enrichment of regions mapping to histone modification peaks in recent DMS. (**c**) Odds ratios measuring the enrichment of different genomic locations in the set of historical DMS (4,054 DMS) for all DMS (in red), DMS more highly methylated in f-AGR with respect to w-RHG (f-AGR > w-RHG, in orange) and DMS more highly methylated in w-RHG with respect to f-AGR (w-RHG > f-AGR, in dark red). (**d**) Odds ratios measuring the enrichment of regions mapping to histone modification peaks in historical DMS. In **a-d**, diamonds represent significant odds ratios (P <0.01, χ^2-test).



Supplementary Fig. 8. Cis location of all meQTLs. Histogram representing the density of detected cis meQTLs as a function of the physical distance from their associated methylation site. 82% of the detected meQTLs were located within 20 kb of the methylation site.



Supplementary Fig. 9. Correlation of meQTL association R2 between RHG and AGR populations. (**a**) Association *R*2 for meQTLs in w-RHG or w-AGR. (**b**) Association *R*2 for meQTLs in e-RHG or e-AGR. (**c**) Association *R*2 for meQTLs in w-RHG or f-AGR. (**a-c**) For each pair of population, *R*2 of associations between SNPs and methylation levels are plotted. meQTLs must be significant at a FDR of 1% either in the RHG or in the AGR population, and with a MAF > 0.1 in both populations. Pearson's correlation coefficients of association *R*2 are indicated on the top of each panel.



Supplementary Fig. 10. **Examples of meQTL-DMS.** (**a-c**) Examples of meQTLs detected in all populations but presenting different allelic frequencies in RHG and AGR groups. The RHG-AGR mean F_{ST} of rs4615933 in *DOCK1*, rs384490 in *CARS* and rs240339 in *LARGE* were higher (F_{ST} 0.06, 0.16 and 0.08, respectively) than that observed genome-wide (RHG-AGR mean F_{ST} < 0.03). (**d**) Example of an meQTL specific to RHG groups. The SNP rs1534362 is associated with differences in methylation in the enhancer region in 6p12.3 (cg23053977) exclusively in RHG. (**a-d**) The three plots on the left represent the distribution of M-values as a function of the genotype, for each population. The minor allele frequency of each meQTL is presented for each population. Red lines indicate the fitted model of a linear regression for M-value ~ genotype for each population. The forest plot on the right represents the estimated β, i.e., the slope of the linear regression.

Supplementary Table 1. Enrichment analysis of *historical* and *recent* DMS in TFBS

	TF ID	*P*-value*	SE†	OR§	Biological functions¶
Historical	TFAP2A	1,73E-08	0,08	1,42	Embryonic cranial skeleton morphogenesis Kidney development Optic vesicle morphogenesis
	GATA2	2,01E-05	0,05	1,32	Urogenital system development Inner ear morphogenesis Negative regulation of neural precursor cell proliferation Positive regulation of erythrocyte differentiation Negative regulation of fat cell differentiation Embryonic placenta development GABAergic neuron differentiation Commitment of neuronal cell to specific neuron type in forebrain
Recent	ELF5	1,88E-30	0,06	1,73	Cell differentiation Mammary gland epithelial cell differentiation
	FOXF2	5,44E-22	0,05	1,61	Epithelial to mesenchymal transition Genitalia development
	NFIL3	1,53E-13	0,05	1,47	Immune response
	FOXC1	3,16E-10	0,05	1,39	regulator of cell viability and resistance to oxidative stress in the eye (UniProtKB/Swiss-Prot)
	NR1H2::RXRA	2,15E-09	0,06	1,38	Regulation of cholesterol homeostasis (UniProtKB/Swiss-Prot)
	FOXI1	9,46E-09	0,05	1,36	Embryo development Mitochondrion distribution Palate development Muscle cell fate determination Regulation of megakaryocyte differentiation Cardiac conduction
	MEF2A	3,87E-08	0,05	1,34	Positive regulation of alkaline phosphatase activity
	MIZF	7,22E-08	0,06	1,34	In utero embryonic development Myoblast differentiation
	ELK4	7,67E-08	0,06	1,34	Cell differentiation
	HNF1A	2,85E-07	0,05	1,32	Glucose homeostasis
	MYC::MAX	1,87E-05	0,06	1,27	Cell proliferation (Entrez) Apoptosis (Entrez)
	PPARG	1,88E-05	0,06	1,27	Cellular response to lithium ion Response to diuretic Organ regeneration Epithelial cell differentiation Cellular response to prostaglandin E stimulus Response to immobilization stress Response to vitamin A
	REL	3,22E-05	0,07	1,26	Heterodimer with NFKB1(UniProtKB/Swiss-Prot)
	FOXA1	4,66E-05	0,05	1,25	Positive regulation of cell-cell adhesion mediated

				by cadherin Positive regulation of mitotic cell cycle Chromatin remodeling Positive regulation of intracellular estrogen receptor signaling pathway Negative regulation of epithelial to mesenchymal transition
	SPIB	5,47E-05	0,06	1,25	Cell differentiation (development of plasmacytoid dendritic cells, UniProtKB/Swiss-Prot)
	TAL1::TCF3	7,55E-04	0,06	1,21	Positive regulator of erythroid differentiation (UniProtKB/Swiss-Prot)
	IRF1	4,01E-03	0,05	1,18	Regulation of MyD88-dependent toll-like receptor signaling pathway Interferon-gamma-mediated signaling pathway Regulation of CD8-positive Alpha-beta T cell proliferation Positive regulation of interleukin-12 biosynthetic process Negative regulation of regulatory T cell differentiation

*P-values were obtained using a chi-squared test to test for enrichment among DMS of high affinity sequences for each transcription factor binding site; [†]Standard errors were obtained using a logistic regression; [§]Odds Ratio measuring the enrichment in high affinity sequences among DMS; [¶]Biological functions refer to Gene Ontology categories other than those directly related to the transcription factor function (such as sequence-specific DNA-binding), unless otherwise mentioned.

Supplementary Table 2. Over-representation analyses of Gene Ontology categories among *recent* differentially methylated genes (total of 3,570)

Category Name	Accession N°	Adj. *P*
Ontology biological process		
immune system process	GO:0002376	8,85E-07
immune response	GO:0006955	1,29E-05
intracellular transport	GO:0046907	2,12E-04
symbiosis, encompassing mutualism through parasitism	GO:0044403	2,63E-04
interspecies interaction between organisms	GO:0044419	2,63E-04
multi-organism cellular process	GO:0044764	4,38E-04
viral process	GO:0016032	4,78E-04
macromolecule localization	GO:0033036	4,90E-04
positive regulation of immune response	GO:0050778	1,06E-03
cellular protein metabolic process	GO:0044267	1,15E-03
protein metabolic process	GO:0019538	1,26E-03
protein transport	GO:0015031	1,60E-03
immune response-activating cell surface receptor signaling pathway	GO:0002429	2,60E-03
positive regulation of immune system process	GO:0002684	2,72E-03
activation of immune response	GO:0002253	2,81E-03
immune response-activating signal transduction	GO:0002757	2,98E-03
single-organism intracellular transport	GO:1902582	2,98E-03
establishment of protein localization	GO:0045184	3,51E-03
regulation of immune response	GO:0050776	3,99E-03
organic substance transport	GO:0071702	4,40E-03
antigen receptor-mediated signaling pathway	GO:0050851	4,40E-03
cellular localization	GO:0051641	4,40E-03
mRNA metabolic process	GO:0016071	4,41E-03
establishment of localization in cell	GO:0051649	5,16E-03
intracellular protein transport	GO:0006886	5,64E-03
protein localization	GO:0008104	5,95E-03
response to stress	GO:0006950	7,93E-03
regulation of immune system process	GO:0002682	7,93E-03
defense response to other organism	GO:0098542	8,24E-03
response to wounding	GO:0009611	1,23E-02
cytokine production	GO:0001816	1,38E-02
translational termination	GO:0006415	1,75E-02
intracellular signal transduction	GO:0035556	1,75E-02
T cell receptor signaling pathway	GO:0050852	1,86E-02
defense response	GO:0006952	1,95E-02
mRNA catabolic process	GO:0006402	1,96E-02
organelle organization	GO:0006996	2,20E-02
myeloid leukocyte activation	GO:0002274	2,24E-02
RNA catabolic process	GO:0006401	2,28E-02
cell activation	GO:0001775	2,33E-02
nuclear-transcribed mRNA catabolic process	GO:0000956	2,56E-02
protein secretion	GO:0009306	2,96E-02
immune effector process	GO:0002252	3,10E-02
cellular macromolecule localization	GO:0070727	3,26E-02
regulation of response to stimulus	GO:0048583	3,27E-02
single-organism organelle organization	GO:1902589	3,61E-02
translational elongation	GO:0006414	3,61E-02
protein targeting	GO:0006605	3,85E-02

cellular protein localization	GO:0034613	3,88E-02
wound healing	GO:0042060	4,50E-02
cellular response to stimulus	GO:0051716	4,58E-02
lymphocyte costimulation	GO:0031294	4,70E-02
Ontology molecular function		
protein binding	GO:0005515	8,95E-07
RNA binding	GO:0003723	1,30E-04
poly(A) RNA binding	GO:0044822	7,44E-04
small molecule binding	GO:0036094	7,44E-04
nucleotide binding	GO:0000166	1,06E-03
nucleoside phosphate binding	GO:1901265	1,07E-03
anion binding	GO:0043168	2,40E-03
ribonucleotide binding	GO:0032553	3,91E-03
nucleoside binding	GO:0001882	3,91E-03
ribonucleoside binding	GO:0032549	4,40E-03
purine ribonucleoside binding	GO:0032550	4,40E-03
purine ribonucleoside triphosphate binding	GO:0035639	4,59E-03
purine nucleoside binding	GO:0001883	4,63E-03
purine ribonucleotide binding	GO:0032555	5,05E-03
purine nucleotide binding	GO:0017076	7,89E-03
catalytic activity	GO:0003824	1,60E-02
transferase activity	GO:0016740	1,75E-02
carbohydrate derivative binding	GO:0097367	2,24E-02
heterocyclic compound binding	GO:1901363	3,31E-02
organic cyclic compound binding	GO:0097159	3,68E-02
Ontology cellular component		
cytoplasm	GO:0005737	6,28E-12
intracellular	GO:0005622	2,09E-09
cytoplasmic part	GO:0044444	2,09E-09
intracellular organelle part	GO:0044446	2,28E-09
intracellular part	GO:0044424	2,28E-09
organelle part	GO:0044422	2,56E-09
intracellular organelle	GO:0043229	1,09E-08
organelle	GO:0043226	4,12E-08
intracellular membrane-bounded organelle	GO:0043231	8,95E-07
membrane-bounded organelle	GO:0043227	2,51E-06
nucleoplasm part	GO:0044451	1,01E-04
vesicle	GO:0031982	1,99E-04
nucleoplasm	GO:0005654	7,87E-04
cytosol	GO:0005829	8,95E-04
extracellular organelle	GO:0043230	1,06E-03
extracellular membrane-bounded organelle	GO:0065010	1,06E-03
extracellular vesicular exosome	GO:0070062	1,06E-03
non-membrane-bounded organelle	GO:0043228	1,06E-03
intracellular non-membrane-bounded organelle	GO:0043232	1,06E-03
membrane-bounded vesicle	GO:0031988	2,98E-03
nuclear lumen	GO:0031981	3,99E-03
intracellular organelle lumen	GO:0070013	4,23E-03
membrane-enclosed lumen	GO:0031974	4,77E-03
organelle lumen	GO:0043233	4,84E-03
vacuole	GO:0005773	7,12E-03
macromolecular complex	GO:0032991	8,16E-03
nuclear part	GO:0044428	8,68E-03
cell part	GO:0044464	1,20E-02

cell	GO:0005623	1,23E-02
ruffle	GO:0001726	1,95E-02
lytic vacuole	GO:0000323	2,20E-02
lysosome	GO:0005764	2,20E-02
nucleus	GO:0005634	2,81E-02
organelle membrane	GO:0031090	3,06E-02
endosome	GO:0005768	3,26E-02

Supplementary Table 3. Over-representation analyses of Gene Ontology categories among *historical* differentially methylated genes (total of 2,130)

Category Name	Accession N°	Adj. *P*
Ontology biological process		
single-multicellular organism process	GO:0044707	2,22E-04
multicellular organismal development	GO:0007275	2,22E-04
multicellular organismal process	GO:0032501	5,45E-04
developmental process	GO:0032502	5,45E-04
cell fate commitment	GO:0045165	6,21E-04
system development	GO:0048731	6,21E-04
single-organism developmental process	GO:0044767	1,44E-03
nervous system development	GO:0007399	1,93E-03
central nervous system development	GO:0007417	5,46E-03
anatomical structure development	GO:0048856	1,00E-02
neuron differentiation	GO:0030182	1,44E-02
anatomical structure morphogenesis	GO:0009653	1,74E-02
tissue development	GO:0009888	1,89E-02
organ morphogenesis	GO:0009887	1,89E-02
generation of neurons	GO:0048699	2,10E-02
signaling	GO:0023052	2,10E-02
single organism signaling	GO:0044700	2,10E-02
locomotion	GO:0040011	2,17E-02
cell motility	GO:0048870	2,17E-02
synaptic transmission	GO:0007268	2,17E-02
cell migration	GO:0016477	2,53E-02
cell communication	GO:0007154	2,76E-02
cell differentiation	GO:0030154	2,79E-02
neurogenesis	GO:0022008	3,42E-02
pattern specification process	GO:0007389	4,03E-02
organ development	GO:0048513	4,09E-02
neuron fate commitment	GO:0048663	4,78E-02
cell-cell signaling	GO:0007267	4,88E-02
Ontology molecular function		
sequence-specific DNA binding	GO:0043565	5,45E-04
nucleic acid binding transcription factor activity	GO:0001071	7,57E-04
sequence-specific DNA binding transcription factor activity	GO:0003700	1,44E-03
transcription regulatory region sequence-specific DNA binding	GO:0000976	1,89E-02
growth factor binding	GO:0019838	2,19E-02
Ontology cellular component		
integral component of plasma membrane	GO:0005887	5,45E-04
intrinsic component of plasma membrane	GO:0031226	1,44E-03
plasma membrane part	GO:0044459	5,23E-03
integral component of membrane	GO:0016021	9,29E-03
intrinsic component of membrane	GO:0031224	2,53E-02

Supplementary Table 4. Over-representation analyses of Gene Ontology categories among *historical* common west-east differentially methylated genes (total of 699)

Category Name	Accession N°	Adj. *P*
Ontology biological process		
single-multicellular organism process	GO:0044707	6,62E-15
multicellular organismal development	GO:0007275	1,18E-14
nervous system development	GO:0007399	1,18E-14
multicellular organismal process	GO:0032501	1,99E-14
developmental process	GO:0032502	4,96E-13
single-organism developmental process	GO:0044767	8,45E-13
system development	GO:0048731	1,03E-12
neuron differentiation	GO:0030182	1,28E-12
neurogenesis	GO:0022008	2,03E-11
generation of neurons	GO:0048699	2,41E-11
anatomical structure development	GO:0048856	7,69E-11
central nervous system development	GO:0007417	7,33E-10
organ development	GO:0048513	3,01E-09
cell differentiation	GO:0030154	3,66E-09
central nervous system neuron differentiation	GO:0021953	1,26E-08
cell development	GO:0048468	1,76E-08
cell fate commitment	GO:0045165	2,25E-08
cellular developmental process	GO:0048869	4,17E-08
tissue development	GO:0009888	1,68E-07
organ morphogenesis	GO:0009887	2,32E-07
anatomical structure morphogenesis	GO:0009653	2,60E-07
regulation of cell differentiation	GO:0045595	2,60E-07
brain development	GO:0007420	2,60E-07
regulation of neuron differentiation	GO:0045664	2,75E-07
cell-cell signaling	GO:0007267	3,27E-07
neuron fate commitment	GO:0048663	7,03E-07
regulation of multicellular organismal process	GO:0051239	8,60E-07
regulation of multicellular organismal development	GO:2000026	1,68E-06
positive regulation of developmental process	GO:0051094	4,53E-06
regulation of developmental process	GO:0050793	5,11E-06
embryo development	GO:0009790	5,72E-06
regulation of neurogenesis	GO:0050767	7,43E-06
positive regulation of cell differentiation	GO:0045597	1,47E-05
regulation of nervous system development	GO:0051960	1,52E-05
pattern specification process	GO:0007389	2,25E-05
muscle structure development	GO:0061061	4,47E-05
single-organism process	GO:0044699	4,47E-05
cell-cell adhesion	GO:0016337	4,47E-05
forebrain development	GO:0030900	4,54E-05
regulation of cell development	GO:0060284	4,54E-05
neuron development	GO:0048666	5,49E-05
synaptic transmission	GO:0007268	7,67E-05
spinal cord development	GO:0021510	8,63E-05

cell differentiation in spinal cord	GO:0021515	9,01E-05
neuron projection development	GO:0031175	1,05E-04
cell morphogenesis involved in differentiation	GO:0000904	1,17E-04
neuron projection morphogenesis	GO:0048812	1,21E-04
axonogenesis	GO:0007409	1,82E-04
epithelium development	GO:0060429	1,86E-04
ventral spinal cord development	GO:0021517	3,09E-04
tissue morphogenesis	GO:0048729	3,79E-04
forebrain neuron differentiation	GO:0021879	5,02E-04
cell morphogenesis involved in neuron differentiation	GO:0048667	5,17E-04
single-organism cellular process	GO:0044763	5,17E-04
cell adhesion	GO:0007155	6,25E-04
forebrain generation of neurons	GO:0021872	6,73E-04
biological adhesion	GO:0022610	6,80E-04
dorsal/ventral pattern formation	GO:0009953	7,19E-04
positive regulation of multicellular organismal process	GO:0051240	8,90E-04
positive regulation of neuron differentiation	GO:0045666	8,90E-04
regulation of calcium ion-dependent exocytosis	GO:0017158	1,14E-03
pallium development	GO:0021543	1,18E-03
axon development	GO:0061564	1,22E-03
cell surface receptor signaling pathway	GO:0007166	1,27E-03
regulation of secretion	GO:0051046	1,29E-03
neuron fate specification	GO:0048665	1,33E-03
tube development	GO:0035295	1,36E-03
embryonic morphogenesis	GO:0048598	1,43E-03
telencephalon development	GO:0021537	1,43E-03
cerebral cortex development	GO:0021987	1,79E-03
behavior	GO:0007610	2,03E-03
neuron migration	GO:0001764	2,28E-03
morphogenesis of an epithelium	GO:0002009	2,37E-03
neurotransmitter transport	GO:0006836	2,65E-03
regulation of neurotransmitter levels	GO:0001505	2,65E-03
locomotory behavior	GO:0007626	2,73E-03
homophilic cell adhesion	GO:0007156	2,92E-03
cell fate specification	GO:0001708	2,99E-03
regionalization	GO:0003002	3,12E-03
response to organic cyclic compound	GO:0014070	3,26E-03
locomotion	GO:0040011	3,87E-03
regulation of exocytosis	GO:0017157	3,98E-03
muscle tissue development	GO:0060537	4,09E-03
positive regulation of secretion	GO:0051047	4,24E-03
sensory organ development	GO:0007423	4,31E-03
striated muscle tissue development	GO:0014706	4,48E-03
negative regulation of neuron differentiation	GO:0045665	4,48E-03
axon guidance	GO:0007411	4,48E-03
neuron projection guidance	GO:0097485	4,48E-03
cell projection morphogenesis	GO:0048858	4,57E-03
response to endogenous stimulus	GO:0009719	4,57E-03
neuron-neuron synaptic transmission	GO:0007270	4,59E-03

response to steroid hormone	GO:0048545	4,84E-03
digestive system development	GO:0055123	4,98E-03
positive regulation of calcium ion-dependent exocytosis	GO:0045956	5,65E-03
anatomical structure formation involved in morphogenesis	GO:0048646	5,90E-03
calcium ion-dependent exocytosis	GO:0017156	6,20E-03
cell part morphogenesis	GO:0032990	6,53E-03
cell migration	GO:0016477	6,56E-03
cellular component movement	GO:0006928	6,74E-03
response to chemical	GO:0042221	8,15E-03
localization of cell	GO:0051674	8,55E-03
signaling	GO:0023052	9,00E-03
single organism signaling	GO:0044700	9,00E-03
monoamine transport	GO:0015844	9,96E-03
muscle cell differentiation	GO:0042692	1,01E-02
regulation of amine transport	GO:0051952	1,05E-02
single-organism behavior	GO:0044708	1,06E-02
response to external stimulus	GO:0009605	1,06E-02
neurotransmitter secretion	GO:0007269	1,12E-02
hindbrain development	GO:0030902	1,12E-02
signal release	GO:0023061	1,18E-02
gland development	GO:0048732	1,18E-02
response to lipid	GO:0033993	1,20E-02
neurological system process	GO:0050877	1,23E-02
negative regulation of developmental process	GO:0051093	1,23E-02
lung development	GO:0030324	1,25E-02
cell communication	GO:0007154	1,27E-02
cellular component morphogenesis	GO:0032989	1,29E-02
cell morphogenesis	GO:0000902	1,32E-02
embryonic organ development	GO:0048568	1,42E-02
chemotaxis	GO:0006935	1,42E-02
taxis	GO:0042330	1,42E-02
G-protein coupled receptor signaling pathway	GO:0007186	1,45E-02
respiratory tube development	GO:0030323	1,46E-02
dopamine transport	GO:0015872	1,47E-02
cell projection organization	GO:0030030	1,48E-02
digestive tract development	GO:0048565	1,75E-02
spinal cord motor neuron differentiation	GO:0021522	1,83E-02
response to cocaine	GO:0042220	1,88E-02
nitric oxide mediated signal transduction	GO:0007263	1,94E-02
amine transport	GO:0015837	1,98E-02
cell motility	GO:0048870	2,06E-02
multicellular organismal response to stress	GO:0033555	2,16E-02
response to alcohol	GO:0097305	2,43E-02
pituitary gland development	GO:0021983	2,52E-02
positive regulation of exocytosis	GO:0045921	2,52E-02
neuropeptide signaling pathway	GO:0007218	2,76E-02
secretion	GO:0046903	2,80E-02
regulation of stem cell proliferation	GO:0072091	2,83E-02
organ formation	GO:0048645	2,85E-02

stem cell differentiation	GO:0048863	3,27E-02
striated muscle cell differentiation	GO:0051146	3,28E-02
negative regulation of cell differentiation	GO:0045596	3,46E-02
secretion by cell	GO:0032940	3,46E-02
respiratory system development	GO:0060541	3,46E-02
exocrine system development	GO:0035272	3,46E-02
muscle organ development	GO:0007517	3,46E-02
developmental induction	GO:0031128	3,68E-02
cellular response to endogenous stimulus	GO:0071495	3,74E-02
neurotransmitter uptake	GO:0001504	3,77E-02
developmental growth	GO:0048589	3,80E-02
synaptic transmission, glutamatergic	GO:0035249	4,13E-02
cell-cell signaling involved in cell fate commitment	GO:0045168	4,30E-02
epithelial tube branching involved in lung morphogenesis	GO:0060441	4,37E-02
regulation of postsynaptic membrane potential	GO:0060078	4,38E-02
regulation of transport	GO:0051049	4,78E-02
regulation of localization	GO:0032879	4,92E-02
Ontology molecular function		
sequence-specific DNA binding	GO:0043565	3,25E-09
nucleic acid binding transcription factor activity	GO:0001071	8,80E-07
sequence-specific DNA binding transcription factor activity	GO:0003700	1,65E-06
transmembrane signaling receptor activity	GO:0004888	1,52E-05
signaling receptor activity	GO:0038023	4,47E-05
receptor activity	GO:0004872	1,48E-04
calcium ion binding	GO:0005509	7,02E-04
signal transducer activity	GO:0004871	8,89E-04
molecular transducer activity	GO:0060089	8,89E-04
transcription regulatory region sequence-specific DNA binding	GO:0000976	2,09E-03
transcription regulatory region DNA binding	GO:0044212	4,07E-03
regulatory region DNA binding	GO:0000975	5,90E-03
regulatory region nucleic acid binding	GO:0001067	5,90E-03
G-protein coupled receptor activity	GO:0004930	6,27E-03
transmembrane receptor protein kinase activity	GO:0019199	3,80E-02
growth factor binding	GO:0019838	3,91E-02
Ontology cellular component		
integral component of plasma membrane	GO:0005887	2,41E-11
intrinsic component of plasma membrane	GO:0031226	6,42E-11
plasma membrane part	GO:0044459	4,17E-08
integral component of membrane	GO:0016021	6,67E-06
intrinsic component of membrane	GO:0031224	1,20E-05
cell periphery	GO:0071944	4,47E-05
plasma membrane	GO:0005886	5,15E-05
membrane part	GO:0044425	8,14E-04
ion channel complex	GO:0034702	2,78E-02
synaptic vesicle	GO:0008021	3,74E-02

Supplementary Table 5. Estimation of the weight of each configuration genome-wide, as given by eQtlBma

Configuration	Weight
e-AGR	6,52E-16
e-RHG	2,33E-01
f-AGR	6,19E-18
w-AGR	2,01E-07
w-RHG	1,21E-01
e-AGR + e-RHG	3,97E-07
e-AGR + f-AGR	9,38E-20
e-AGR + w-AGR	1,03E-10
e-AGR + w-RHG	4,62E-10
e-RHG + f-AGR	2,77E-13
e-RHG + w-AGR	4,96E-04
e-RHG + w-RHG	2,11E-02
f-AGR + w-AGR	3,18E-11
f-AGR + w-RHG	6,87E-04
w-AGR + w-RHG	6,71E-03
e-AGR + e-RHG + f-AGR	4,90E-16
e-AGR + e-RHG + w-AGR	6,39E-04
e-AGR + e-RHG + w-RHG	1,40E-08
e-AGR + f-AGR + w-AGR	3,35E-03
e-AGR + f-AGR + w-RHG	6,76E-09
e-AGR + w-AGR + w-RHG	1,62E-06
e-RHG + f-AGR + w-AGR	7,45E-12
e-RHG + f-AGR + w-RHG	8,86E-05
e-RHG + w-AGR + w-RHG	1,23E-04
f-AGR + w-AGR + w-RHG	8,52E-03
e-AGR + e-RHG + f-AGR + w-AGR	1,48E-02
e-AGR + e-RHG + f-AGR + w-RHG	1,25E-16
e-AGR + e-RHG + w-AGR + w-RHG	1,67E-10
e-AGR + f-AGR + w-AGR + w-RHG	4,50E-02
e-RHG + f-AGR + w-AGR + w-RHG	6,93E-05
e-AGR + e-RHG + f-AGR + w-AGR + w-RHG	5,44E-01

Supplementary Table 6. Primer sequences used for bisulfite PCR-pyrosequencing

Enhancer_cg23053977 F: GGG GTA TAG TAG AAG AAA ATT TTA AGG Enhancer_cg23053977 R: /5BiodT/TT TTC CCA CAA TAC AAC TAA ATA TTC AC Enhancer_cg23053977 S: TTT TTA GAA TAA AGT AAG TAT TTA ATG AT
ADAM28_ cg18757155 F: GAG TAT GGT AAA GGA GAG TAA AAT AAT AGT ADAM28_ cg18757155 R: /5BiodT/AA ATA AAA CTC CCT AAT TCA TTC TTA TCT ADAM28_ cg18757155 S: GGA GGT AGT TAG GAT TT
RORA_ cg09879458 F: GAT TAT ATT TTG TGG GGT GAA TGG AGG TA RORA_ cg09879458 R: /5BiodT/CC TCC CAA CCT TTA TTA TTC CTT TTC C RORA_ cg09879458 S: AGT AAT ATA TAG TAG TAT GAG AAA T
COL23A1_ cg08684511 F: GGT GTT TGT AGT TTA AGG GTA TGT AG COL23A1_ cg08684511 R: /5BiodT/CA ACT AAA AAC TAA CAC CAT ATA CCT COL23A1_ cg08684511 S: TGA GGA GTG AAA TTG TAT TTA ATT AT

Supplementary Notes

Supplementary Note 1. Different Methylation Profiles across Genomic Regions.
Individual methylation data showed the expected bimodal profiles (Supplementary Fig. 2a), with M-values ranging from -6 to +6 and two peaks at -3.5 and +3.5, corresponding to unmethylated sites and fully methylated sites, respectively. The level of methylation decreased near and within CpG islands (Supplementary Fig. 2b). We observed a lower level of methylation at sites around the transcription start sites of genes than in intergenic and gene body regions (Supplementary Fig. 2c). In addition, sites in gene bodies were slightly more methylated than intergenic sites, as previously reported [24]. Finally, when looking at inter-individual variation in methylation levels, we observed that the correlation between methylation profiles of technical replicates (19 replicates in total) was higher than the correlation between randomly chosen pairs of samples from the same population (Supplementary Fig. 2d). With respect to global methylation profiles, samples from w-AGR and e-AGR populations, which showed stronger correlations between pairs of samples, tended to be more homogeneous than samples from f-AGR, w-RHG and e-RHG.

Supplementary Note 2. Age Imputation from Methylation Data.
As age has been shown to impact on the variation of DNA methylation[61], we controlled for this potential bias in our DMS and meQTL analyses. Ages were available for all samples from the Gabon/Cameroon region (w-RHG, w-AGR and f-AGR), but not for those from Uganda (e-RHG and e-AGR). We thus imputed ages for all samples, using an elastic net regression method[51], and compared predicted with declared ages, when the latter were available. Although predicted ages were, on average, ~7 years older than declared ages, we observed a high correlation between the two, with a Pearson correlation coefficient of 0.84 (Supplementary Fig. 3). Given the relative accuracy of the predictive model, and for the sake of consistency between comparisons, we used the predicted ages for all samples as a covariate in the linear regressions performed for the DMS and meQTL analyses.

Supplementary Note 3. Accounting for heterogeneity in blood cell composition in methylation analyses.
Because environmental factors could alter the relative abundance of different cell types in the blood of the populations studied, and thus confound our analysis[62], we predicted the proportions of different cells types in unfractionated whole blood, as previously described[50], across all 352 samples. We found that our predictions fall within generally accepted ranges: 47% for granulocytes, 11% for $CD8^+$-T cells, 14% for $CD4^+$-T cells, 12% for NK cells, 10% for B cells and 6% for monocytes. To check for the accuracy of the predictive model, we compared these estimations with fluorescence-activated cell sorting (FACS) data for peripheral blood mononuclear cells (PBMCs) from 35 e-RHG and 31 e-

AGR. Our results showed high correlation coefficients (Pearson's *R*: 0.48-0.57) between estimated and observed proportions for the cellular types from which we could obtain FACS data ($CD4^+$ T-cells, $CD8^+$ T-cells, B-cells, and NK-cells) (Supplementary Fig. 4). In light of this, we used the estimated cell subtype proportions to adjust M-values for subsequent analyses (i.e., PCA, DMS, and meQTL mapping). Finally, we assessed the efficiency of this correction by examining the extent to which our PCA were correlated with variation in blood cell composition. While before the correction by age, sex and predicted cell types proportions, PC1 correlated significantly with T cells ($P = 4.1\times10^{-3}$) and B cells ($P = 6.1\times10^{-3}$), these correlations became non significant after correction.

Supplementary Note 4. Population specificity of meQTLs in whole blood. Fraser and colleagues[22] compared CEU and YRI populations from the HapMap Project, using the HumanMethylation27 BeadChip assay on lymphoblastoid cell lines (LCLs). They found that only 8.9% of detected meQTLs were shared between the two populations, consistent with the extensive population specificity of DNA methylation heritability estimates. Moen and colleagues[24] compared the same populations and cell lines, but used the HumanMethylation450 BeadChip assay. To illustrate the extent of population specificity of their meQTLs, they reported a very low correlation (−0.03) between SNP association R^2 estimated in CEU and YRI, suggesting that genetic epistasis and/or G×E interactions could be common in explaining DNA methylation variation[24]. In our study, we estimated that 90% of meQTLs are shared across populations. To test whether this proportion could be accounted for by a bias of eQtlBma against population-specific meQTL associations, we reanalyzed our data using the approach of Moen and colleagues[24]. We estimated R^2 coefficients of a linear regression model of DNA methylation on SNP genotypes, correcting for age, gender, ancestry and immune cell proportions. We obtained strong correlations (0.39 to 0.66) between association R^2 estimated in RHG and AGR populations (Supplementary Fig. 9). This result attests that eQtlBma is not biased against population-specific meQTL associations, and that most meQTLs detected here are indeed shared among populations. We suggest that this pattern result from (i) the lower degree of genetic differentiation of our populations (F_{ST}=0.03-0.05)[7] with respect to CEU and YRI (F_{ST}=0.12)[59], and (ii) DNA methylation was assessed here in whole blood, a complex mixture of different cell types, while the studies mentioned above used LCLs.

61 Bell, J. T. *et al.* Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet* **8**, e1002629 (2012).

62 Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol* **15**, R31 (2014).

Maud FAGNY

Sujet : L'Homme face à son environnement: une histoire génétique et épigénétique du génome humain

Résumé : Les populations humaines ont été confrontées à de nombreux changements environnementaux au cours de leur histoire et présentent aujourd'hui une grande diversité d'habitats et de modes de subsistance. Cependant, l'ampleur de l'adaptation génétique et des réponses épigénétiques à ces changements est débattue. Nous avons d'abord étudié la puissance de diverses statistiques pour détecter les balayages sélectifs dans le contexte des données de séquençage à haut débit, et évalué leur robustesse à différents facteurs confondants. En utilisant des jeux de données de séquençage, nous montrons que les balayages sélectifs ont eu un impact modéré mais non négligeable dans l'évolution récente du génome humain. Les régions sous sélection sont enrichies en mutations associées à des variations phénotypiques. Nous avons ensuite évalué l'impact respectif des facteurs génétiques et environnementaux sur la diversité épigénétique humaine. Pour cela, nous avons obtenu les génotypes et les profiles de méthylation de l'ADN de populations d'Afrique Centrale présentant des différences récentes d'habitat ou historiques de modes de vie et de profil génétique. Nous montrons que les deux facteurs ont un effet similaire sur le méthylome mais diffèrent par les fonctions biologiques affectées et les mécanismes expliquant les variations observées. Plus généralement, les variations de méthylation sont fortement associées à des mutations génétiques qui sont enrichies en signaux de sélection positive. En conclusion, ce travail apporte un aperçu de la contribution des mutations génétiques et des réponses épigénétiques à l'adaptation humaine aux changements environnementaux sur plusieurs échelles de temps.

Subject : Humans in an adaptive world: genetic and epigenetic responses to environmental challenges

Abstract : Human populations have faced a large number of environmental challenges during their evolutionary history and present today a wide range of habitats and mode of subsistence. However, the extent of genetic adaptation and epigenetic responses to such environmental variation remains controversial. We first explored the power of several statistics to detect hard selective sweeps in the context of whole-genome sequencing data, and evaluated their robustness to demography and other selection modes. Using data from the 1,000 Genomes Project and Complete Genomics, we showed that hard sweeps targeting low-frequency standing variation have played a moderate, albeit significant, role in recent human evolution. The signals of selection detected were moreover enriched in functional variants detected by genome-wide association studies. We then evaluated the relative impacts of genetic and environmental factors on human epigenomic diversity. To do so, we generated genome-wide genetic and DNA methylation profiles for Central African populations differing in their current habitat or in their historical lifestyle and genetic background. We found that both factors have similar critical impacts on the shaping of the global methylome, but the biological functions affected and the mechanisms underlying DNA methylation variation strongly differ. More generally, methylation variation shows strong associations with nearby genetic variants that, moreover, are enriched in signals of natural selection. Together, this work provides new insight into the contribution of genetic adaptation and epigenetic responses to the adaptation of humans to environmental changes over different time scales.